



**This electronic thesis or dissertation has been
downloaded from Explore Bristol Research,
<http://research-information.bristol.ac.uk>**

Author:

Zheng, Jie

Title:

Fine Mapping of Genetic Variants Influencing Complex Traits in Human

General rights

Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>. This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy

Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

- Your contact details
- Bibliographic details for the item, including a URL
- An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.



**This electronic thesis or dissertation has been
downloaded from Explore Bristol Research,
<http://research-information.bristol.ac.uk>**

Author:

Zheng, Jie

Title:

Fine Mapping of Genetic Variants Influencing Complex Traits in Human

General rights

Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>. This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy

Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

- Your contact details
- Bibliographic details for the item, including a URL
- An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

University of Bristol

Fine Mapping of Genetic Variants Influencing Complex
Traits in Human

Zheng Jie

A dissertation submitted to the University of Bristol in
accordance with the requirements for award of the degree of
Doctor of philosophy in the Faculty of Medicine and
Dentistry

**SCHOOL OF SOCIAL AND COMMUNITY
MEDICINE**

September 2014

Thesis Abstract

Genome-wide association studies (GWAS) have successfully identified thousands of single nucleotide polymorphisms (SNPs) associated with complex traits and diseases of human. Meta-analysis of multiple GWASs has become a popular method since it increases the statistical power and reduces false positive findings. In the genetic field, fine mapping is a process of identifying independently associated variants within a genomic region, which improves our understanding of the causal mechanisms underlying human diseases. However, data sharing among different studies is usually unavailable for meta-analysis of a large number of studies. Fine mapping of genetic variation under the current meta-analysis system is administratively onerous and time consuming. Therefore, statistical approaches which can process fine mapping analysis using meta-analysis summary statistics are assuming increasing importance.

This thesis is concerned with the development of two fine mapping methods using meta-analysis summary statistics: Sequential Sentinel SNP Regional Association Plots (SSS-RAP) and haplotype-based regional association analysis program (HAPRAP). SSS-RAP detects SNPs with independent effects conditional on the top associated signal using meta-analysis summary statistics and summary pair-wise SNPs haplotype frequencies obtained from reference genotype panel. I demonstrate that SSS-RAP is as powerful as conditional analysis and ten model selection methods in individual-level. I applied SSS-RAP to meta-analysis of Electrocardiography (ECG) traits, gallbladder disease (GBD) traits and GIANT BMI database. In addition, HAPRAP is an empirical EM style approach, which extends multiple regression and conditional analysis to meta-analysis levels. I demonstrate that HAPRAP is statistically outperforming existing methods in meta-analysis level. I applied HAPRAP to meta-analysis for ECG traits. Finally, I discuss the position of SSS-RAP and HAPRAP in genetic fine mapping and future direction of genetic fine mapping.

Acknowledgements

Frist, I would like to thank my PhD supervisors: Prof. Ian Day, Dr. Tom Gaunt and Dr. Santiago Rodriguez. I am especially indebted to Ian and Tom for hours of guidance, advice and support – as well as giving me the freedom and the motivation to pursue and evolve my own ideas. I would like to give my special thank to Prof. David Evans and Dr. Lavinia Paternoster for their support.

I would also like to thank all my friends in the Bristol Genetic Epidemiology Laboratories (BGEL). Special thanks go to Denis Baird, Mesut Erzurumluoğlu, Dr. Hashem Shihab, Tom Richardson, Lea Trela-Larsen and Anna Guyatt for useful discussions and their help in reviewing this work.

I am grateful to my father and mother: Zheng Wenbing and Zhang Hua. For without their encouragement and support over the past 30 years, I would have never been able to pursue a PhD in the first place.

Last, but by no means least, I would like to thank my precious wife: Lin Jiayu. If it wasn't for her unconditional love, patience and support, I would have never made it this far – I am who I am because of them.

Author Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED _____Jie Zheng_____

DATE _____02/02/2015_____

Table of Contents

Chapter 1. Introduction.....	1
1.1 Background of human genome.....	1
1.1.1 Human genome.....	1
1.1.2 Deoxyribonucleic acid (DNA).....	2
1.1.3 Gene.....	2
1.1.4 Gene expression.....	3
1.1.5 Genetic variation.....	4
1.1.6 Genetic disorder.....	8
1.2 Population Genetics concepts.....	12
1.2.1 Allele frequencies.....	12
1.2.2 Disease-variant hypothesis and missing heritability.....	15
1.2.3 Linkage disequilibrium, haplotypes and haplotype phasing.....	19
1.2.4 Genotype Imputation.....	24
1.3 Mapping genetic variants to complex traits.....	29
1.3.1 Phenotypes and phenotype standardization.....	30
1.3.2 Association analysis for candidate genes.....	33
1.3.3 Genome-wide association analysis (GWAS).....	33
1.3.4 Regression analysis.....	25
1.3.5 Meta-analysis of Genome-wide association analysis.....	30
1.3.6 Genetic fine mapping.....	33
1.4 Existing model selection approaches for identifying independent SNPs associated with complex traits.....	44
1.4.1 Stepwise regression and conditional analysis.....	47
1.4.2 Model selection approaches.....	48
1.5 Materials.....	50
1.5.1 British Women's Heart and Health Study (BWHHS).....	50
1.5.2 HapMap and the 1000 Genome Project.....	53
1.5.3 Avon Longitudinal Study of Parents and Children (ALSPAC).....	55

1.5.4	Meta-analysis of cardiovascular disease traits.....	56
1.5.5	Meta-analysis of gallbladder disease.....	57
1.5.6	Meta-analysis of Atopic Dermatitis.....	57
1.5.7	Meta-analysis of Body Mass Index.....	58
1.6	Bioinformatics tools.....	59
1.6.1	Variants database: dbSNP.....	59
1.6.2	SNP Annotation tools: VEP and SNAP.....	60
1.6.3	Batch Coordinate Conversion (LiftOver).....	61
1.6.4	Linkage disequilibrium tools: Haploview, CubeX and MIDAS.....	62
1.6.5	Functional prediction tools: FATHMM and RegulomeDB.....	64
1.7	Statistical genetics tools.....	66
1.7.1	PLINK.....	66
1.7.2	GCTA conditional and joint SNP effect analysis.....	66
1.7.3	R and Python.....	67
1.8	Major aims and objectives.....	68
1.5.1	Objectives.....	68
Chapter 2. Sequential sentinel SNP regional association plots using meta-analysis summary statistics.....		70
2.1	Introduction.....	70
2.1.1	One Limitation of single SNP model GWAS, conditional analysis and model selection approaches.....	71
2.1.2	Meta-analysis and limitation of current meta-analysis system.....	72
2.1.3	Haplotype, haplotype frequency and the r^2 measure of linkage disequilibrium.....	74
2.1.4	Limitation of considering r^2 as the standard of SNP independence.....	74
2.1.5	Aims and Objectives.....	77
2.2	Methods and Materials.....	79
2.2.1	Formulation of problems: linking meta-analysis summary statistics to pair-wise linkage disequilibrium measures.....	81
2.2.2	Relating regression coefficients to minor allele frequencies and haplotype	

frequencies.....	82
2.2.3 Estimating standard errors of transformed betas.....	84
2.2.4 Two tailed t test and stepwise elimination.....	85
2.2.5 SSS-RAP web-based tool.....	88
2.2.6 Sample datasets.....	88
2.3 Results.....	92
2.3.1 Estimating standard errors of transformed betas.....	92
2.3.2 Relationship between r^2 and beta estimation.....	94
2.3.3 Example of a SSS-RAP plot.....	96
2.3.4 Performance comparison of SSS-RAP, model selection and conditional analysis using BWHHS individual level data.....	97
2.3.5 Applying SSS-RAP to an admixed population using 1000 Genome Project data.....	101
2.3.6 Applying SSS-RAP to 210 simulated populations.....	102
2.4 Discussion.....	103
2.4.1 Advantage of SSS-RAP.....	104
2.4.2 Assumptions and limitations of SSS-RAP.....	107
2.4.3 Conclusion.....	110
Chapter 3. Fine mapping of cardiovascular disease related traits and application of SSS-RAP.....	111
3.1 Introduction.....	111
3.1.1 Electrocardiography (ECG).....	112
3.1.2 Body mass index (BMI)	115
3.1.3 Gallbladder disease (GBD).....	117
3.1.4 Aims and objectives.....	118
3.2 Methods and Materials.....	119
3.2.1 Sample datasets.....	119
3.2.2 Applying SSS-RAP to meta-analysis of ECG traits and performance comparison of SSS-RAP in group level.....	120
3.2.3 Applying SSS-RAP to GIANT meta-analysis of BMI.....	121

3.2.4	SSS-RAP analysis of GBD traits and identifying of relationship between GBD associated SNPs and lipids associated SNPs.....	124
3.3	Results.....	126
3.3.1	Applying SSS-RAP program to ECG meta-analysis summary results and four data combinations.....	127
3.3.2	SSS-RAP analysis of GIANT meta-analysis of BMI..	128
3.3.3	SSS-RAP analysis of GBD associated SNPs and analysis of loci associated with lipids traits.....	129
3.4	Discussion.....	137
3.4.1	Comparing the performance of SSS-RAP using different resources...	137
3.4.2	SSS-RAP analysis of ECG traits.....	138
3.4.3	Meta-analysis of GIANT BMI dataset.....	139
3.4.4	Conclusion.....	141
Chapter 4.	Haplotype-based group-level association analysis using an E-M algorithm.....	143
4.1	Introduction.....	143
4.1.1	Multiple regression.....	143
4.1.2	GCTA conditional and joint effect analysis.....	145
4.1.3	Haplotypes and haplotype phasing.....	147
4.1.4	Expectation-Maximization algorithm.....	149
4.1.5	Aims and Objectives.....	151
4.2	Methods and Materials.....	153
4.2.1	Formulation of problems: Linking marginal effect of SNP to independent SNP effect.....	153
4.2.2	Estimating standard error of HAPRAP betas using simulation based HAPRAP.....	158
4.2.3	Flowchart of HAPRAP program.....	161
4.2.4	HAPRAP web-based tool.....	162
4.2.5	Sample datasets.....	163
4.2.6	Performance comparisons.....	167

4.3 Results.....	170
4.3.1 The independent benchmark of HAPRAP, GCTA and multiple regression using simulated meta-analysis data.....	170
4.3.2 A performance comparison of HAPRAP against GCTA using BWHHS genotype data.....	174
4.3.3 A performance comparison of HAPRAP against GCTA using 1000 Genome genotype data.....	178
4.4 Discussion.....	182
4.4.1 Advantages of HAPRAP.....	182
4.4.2 Assumptions and limitations for HAPRAP.....	184
4.4.3 Conclusion.....	185
Chapter 5. Fine mapping of meta-analysis group-level data using HAPRAP.....	187
5.1 Introduction.....	187
5.1.1 Blood lipids traits.....	187
5.1.2 Coagulation.....	194
5.1.3 Other CVD related traits.....	195
5.1.4 Atopic dermatitis.....	197
5.1.5 Aims and objectives.....	198
5.2 Methods and Materials.....	199
5.2.1 Sample datasets.....	199
5.2.2 Fine mapping of GBD outcomes.....	200
5.2.3 Fine mapping of UCLEB meta-analysis of CVD related traits.....	201
5.2.4 Fine mapping of Atopic Dermatitis.....	203
5.3 Results.....	204
5.3.1 Fine mapping of GBD outcomes.....	204
5.3.2 Fine mapping of ECG traits.....	204
5.3.3 Fine mapping of lipids traits.....	205
5.3.4 Fine mapping of coagulation traits.....	205
5.3.5 Fine mapping of other CVD related traits.....	211
5.3.6 Fine mapping of Atopic Dermatitis.....	213

5.4 Discussion.....	215
5.4.1 Fine mapping of ABCG5-8 region.....	215
5.4.2 Fine mapping of NOS1AP region.....	215
5.4.3 Fine mapping of SCN5A-10A region.....	216
5.4.4 Fine mapping of candidate regions associated with Atopic Dermatitis.....	216
5.4.5 Conclusion.....	217
Chapter 6. Conclusions and future direction.....	219
6.1 Selecting independent SNPs using SSSRAP.....	221
6.2 Haplotype based model selection using HAPRAP.....	222
6.3 Future directions.....	223
6.3.1 Further development of SSS-RAP.....	223
6.3.2 Further development of HAPRAP.....	223
6.4 Conclusion.....	227
Chapter 7. Reference.....	228
Chapter 8. Appendices.....	251
8.1 Supplementary Notes.....	251
8.1.1 Codes for LiftOver.....	251
8.1.2 R codes used for model selection methods.....	251
8.1.3 Funding section of the samples.....	252
8.2 Supplementary Tables.....	257
8.3 Supplementary Figures.....	322

LIST OF TABLES

Table 1-1. *Approximate-low frequency / rare variant GWAS platform content. K means thousand and M means million...16*

Table 1-2. *Example of a 2X2 haplotype table...20*

Table 1-3. *Haplotypes and haplotype uncertainty22*

Table 1-4. *An example of meta-analysis summary statistics...432*

Table 1-5. *Example of BED file...61*

Table 1-6. *The RegulomeDB score system...65*

Table 2-1. *Examples of query dependent SNPs with low LD ($r^2 < 0.223$) to the top hits but dependent on the top hits...75*

Table 2-2. *Assumptions and notations...80-81*

Table 2-3. *Possible haplotype frequencies combination for the 3 SNPs model...91-92*

Table 2-4. *Summary information of BWHHS participants. P-value threshold is the cut off value of the SNP significant test...97*

Table 2-5. *Test of one ancestral group assumption for SSS-RAP...102*

Table 2-6. *Computational time of SSS-RAP...106*

Table 3-1. *FTO and MC4R SNPs reported to be associated with BMI in the GIANT meta-analysis...117*

Table 3-2. *Content of combination of different data-sets, X11 represents the haplotype frequencies of two major alleles of pair-wise SNPs...121*

Table 3-3. *Genes that found extra independent effect SNPs by SSS-RAP program in GIANT BMI database...129*

Table 3-4. *Comparison of performance of SSS-RAP in individual-level, meta-analysis level and different reference panels in GBD traits...134*

Table 4-1. *Notation of HAPRAP algorithm...154*

Table 4-2. *List of phenotypes and genotypes selected from BWHHS...163*

Table 4-3. *Assumptions of Pair-Wise LD (r^2) of SNPs Models for Population*

Simulations...167

Table 4-4. *Performance Comparison of HAPRAP, GCTA and Multiple Regression Using a Simulated Population with Two Rare Variants...173*

Table 4-5. *The Results of Concordance Correlation Coefficient Analysis...178*

Table 5-1. *Three major isoforms of APOE gene...192*

Table 5-2. *P value threshold for each atopic dermatitis candidate region...204*

Table 5-3. *HAPRAP and GCTA joint effect analysis of identifying two independent effect SNPs associated with gallbladder disease in ABCG 5 and 8 regions...204*

Table 5-4. *Summary of multiple associated SNPs at NOS1AP for QTc interval and SCN5A-10A for PR interval...206*

Table 5-5. *Independent effect SNPs identified by HAPRAP in four lipids traits...208*

Table 5-6. *Independent effect SNPs identified by HAPRAP in three coagulation traits...210*

Table 5-7. *Independent effect SNPs identified by HAPRAP in crp and bilirubin...212*

Table 5-8. *Genomic regions with multiple SNPs associated with Atopic Dermatitis using fixed effect model meta-analysis results...214*

Table 6-1. *A list of consortium with public available meta-analysis summary data...220*

Table 8-1A. *Significant SNPs for lipids traits in BWHHS individual level data...258*

Table 8-1B. *Significant SNPs for ECG traits in BWHHS individual level data...260*

Table 8-2. *Previous GWAS findings of PR interval...261*

Table 8-3. *Previous GWAS findings of QRS interval...261*

Table 8-4A. *Previous GWAS findings of QTc interval...263*

Table 8-4B. *Previous GWAS findings of QT interval...264*

Table 8-5. *Significant SNPs for ECG traits in meta-analysis...267*

Table 8-6. *The significant SNPs of the GBD meta-analysis...267*

Table 8-7. *Associations between GBD and 63 independent loci previously reported to be associated with lipid levels...270*

Table 8-8. *Performance Comparison of HAPRAP, GCTA and Multiple Regression using Artificial Meta-Analyses of the Simulated Populations...277*

Table 8-9. *Performance Comparison of HAPRAP and GCTA Using BWHHS Summary Statistics and BWHHS Individual-level Genotypes...284*

Table 8-10. *Performance Comparison of HAPRAP and GCTA Using BWHHS Summary Statistics and 1000 Genome Individual-Level Genotypes...288*

Table 8-11. *Genetic variants reported to be associated with HDL-C by previous GWASs...292*

Table 8-12. *Genetic variants reported to be associated with LDL-C by previous GWASs...295*

Table 8-13. *Genetic variants reported to be associated with triglycerides by previous GWAS studies...298*

Table 8-14. *Novel genetic variants reported to be associated with total cholesterol by previous GWASs...298*

Table 8-15. *Genetic variants reported to be associated with FVII by previous GWASs...299*

Table 8-16. *Genetic variants reported to be associated with FVIII by previous GWASs...299*

Table 8-17. *Genetic variants reported to be associated with vWF by previous GWASs...299*

Table 8-18. *Genetic variants reported to be associated with CRP by previous GWASs...301*

Table 8-19. *Genetic variants reported to be associated with Bilirubin by previous GWASs...302*

Table 8-20. *Genetic variants reported to be associated with urate by previous GWASs...*305

Table 8-21. *Genetic variants reported to be associated with atopic dermatitis by previous GWASs...*306

Table 8-22. *SNPs significantly associated with GBD in the GBD meta-analysis...*306

Table 8-23. *394 SNPs selected for fine mapping analysis in the UCLEB meta-analysis of CVD related traits...*321

LIST OF FIGURES

Figure 1-1. *Human Genome...1*

Figure 1-2. *Relationship between allele frequencies of genetic variants and penetrance of diseases/traits...15*

Figure 1-3. *Relationship between power and sample size with different MAF...16*

Figure 1-4. *Comparison of odds ratio between rare and common variants...19*

Figure 1-5. *Example of genotype imputation...26*

Figure 1-6. *Example of scatter plot and linear regression of a continuous explanatory variable...36*

Figure 1-7. *Example of association analysis between a biallelic SNP and a quantitative phenotype...39*

Figure 1-8. *Relationship between linkage analysis, association analysis and fine mapping...43*

Figure 1-9. *Structure of the BWHHS study...52*

Figure 1-10. *Haplotype interface and example of a LD plot...62*

Figure 1-11. *Web-based tool of CubeX...63*

Figure 1-12. *Example of MIDAS results plot...64*

Figure 2-1. *A special case using the three SNPs model...77*

Figure 2-2. *Two SNPs 2x2 haplotype table and regression coefficient transformation...83*

Figure 2-3. *Example of a linear trend line for a single locus...87*

Figure 2-4. *Flowchart of the SSSRAP approach...88*

Figure 2-5. *Homepage of SSS-RAP programme...91*

Figure 2-6. *The three SNPs model for the admixed population simulations...86*

- Figure 2-7.** *The influence of minor allele frequency on the standard error of beta...93*
- Figure 2-8.** *Comparison of the observed standard errors to the simulated standard errors and the transformed standard errors in different r^2 ranges...94*
- Figure 2-9.** *Relationship between r^2 and beta estimate of the possible dependent SNP for a given major allele frequency of the top hit and observed beta for the top SNP...95*
- Figure 2-10.** *Example of a SSS-RAP plot...96*
- Figure 2-11.** *Comparison of model selection approaches, conditional analysis and SSS-RAP using BWHHS individual level data in NOS1AP region...99*
- Figure 2-12.** *Comparison of model selection approaches, conditional analysis and SSS-RAP using BWHHS individual level data in other regions...100*
- Figure 2-13.** *SSS-RAP and conditional analysis results for the three SNPs haplotype simulation...103*
- Figure 2-14.** *An example of one ‘bystander’ efficiently represents two separate causal sites themselves in LD...109*
- Figure 3-1.** *ECG waves and intervals...113*
- Figure 3-2.** *Body mass index chart...116*
- Figure 3-3.** *Predicting standard error using minor allele frequency...122*
- Figure 3-4.** *Flow chart of the beta prediction...123*
- Figure 3-5.** *Comparison between SSS-RAP and conditional analysis using meta-analysis data in ECG traits...127*
- Figure 3-6.** *Comparison between SSSRAP and conditional analysis using meta-analysis data in ECG traits...128*
- Figure 3-7.** *Comparison between r^2 values and P values and between r^2 values and effect sizes for key SNPs associated with GBD...131-133*
- Figure 3-8.** *Pathway analysis relating GBD with SNPs associated with plasma lipid levels...136*
- Figure 4-1.** *Example of haplotypes among four loci for different populations...149*

Figure 4-2. *An overview of the E-M algorithm...150*

Figure 4-3. *Schematic diagram of HAPRAP...154*

Figure 4-4. *The SNPs-haplotypes matrix used by HAPRAP...155*

Figure 4-5. *Population simulation for simHAPRAP...160*

Figure 4-6. *Flowchart for simHAPRAP...161*

Figure 4-7. *Workflow of HAPRAP program...162*

Figure 4-8. *HAPRAP web-based tool...163*

Figure 4-9. *The Multidimensional Scaling Plot of 1000 Genome CEU and GBP Population Together with BWHHS Individuals...165*

Figure 4-10. *Performance Comparison of HAPRAP, GCTA and Multiple Regression Using Meta-Analyses of the Simulated Populations...171-172*

Figure 4-11. *Performance comparison against HAPRAP and GCTA joint effect analysis using BWHHS unphased genotypes...175*

Figure 4-12. *Performance Comparison of HAPRAP and GCTA Joint Effect Analysis Using BWHHS Unphased Genotypes...177*

Figure 4-13. *Performance comparison between HAPRAP and GCTA joint effect analysis using 1000 Genome phased genotypes...179*

Figure 4-14. *Performance Comparison against HAPRAP and GCTA Joint Effect Analysis Using 1000 Genome Phased Genotypes...181*

Figure 5-1. *Cholesterol and lipoprotein transport pathways...189*

Figure 8-1. *Performance comparison of HAPRAP, GCTA and multiple regression using artificial meta-analyses of the simulated populations....325*

LIST OF PUBLICATIONS

- The new method and applications described in Chapter 3 and Chapter 4 has been peer-reviewed and published in *Annul of Human Genetics*:

Jie Zheng, Tom R Gaunt, Ian N M Day. “Sequential sentinel SNP Regional Association Plots (SSS-RAP): an approach for testing independence of SNP association signals using meta-analysis data”. *Annals of Human Genetics*. 2013 January; 77(1):67-79. (<http://www.ncbi.nlm.nih.gov/pubmed?term=sss-rap>)

- The gallbladder disease fine mapping analysis described in Chapter 4 and Chapter 6 has been submitted for publication in *European Journal of Human Genetics*

Santiago Rodriguez[†], Tom R Gaunt[†], Yiran Guo[†], **Jie Zheng**, Weihang Tang, Fazal Danish , Andrew Johnson, Berta A.Castillo, Yun R Li, Hakon Hakonarson, Sarah G Buxbaum, Tom Palmer, Michael Y Tsai, Leslie A Lange, Shah Ebrahim, George Davey Smith, Debbie A Lawlor, Aaron R. Folsom, Ron Hoogeveen, Alex Reiner, Brendan Keating, Ian N M Day “Lipids, obesity and gallbladder disease in women: insights from genetic studies using the cardiovascular gene-centric 50K SNP array”. *Eur J Hum Genet*. 2015 Apr 29.

- The tables and figures described in Chapter 5 and Chapter 6 have been submitted for publication in *Bioinformatics*

Jie Zheng, Santiago Rodriguez, Jon White, Claudia Giambartolomei, Delilah Zabaneh, Richard Morris, Meena Kumari, Juan P. Casas, Aroon D. Hingorani, Tom R. Gaunt[†] Ian N M Day “HAPRAP: a haplotype-based iterative method for genetic fine mapping using complex traits meta-analysis summary results”. *Bioinformatics*. 2015, under revision

- Fine mapping of Atopic Dermatitis meta-analysis is part of the following paper Lavinia Paternoster, ... **Jie Zheng** ...et al, “Multi-ethnic genome-wide association study of 21,000 cases and 95,000 controls identifies 11 novel risk loci for atopic dermatitis”. *Nature Genetics* 2015, under revision

CHAPTER 1. INTRODUCTION

1.1. Background of human genetics

Human genetics is the study of inheritance as it occurs in human beings. Here, I will give a brief overview of the human genetics concepts, including human genome, Deoxyribonucleic Acid (DNA), gene, gene expression, genetic variation and genetic disorder.

1.1.1. Human genome

The human genome is composed of approximately 3 billion base pairs (International Human Genome Sequencing Consortium, 2004). The human genome is composed of 23 pairs of chromosome (Figure 1-1). 22 autosomes and the one pair of sex-determining chromosomes: X/X (female) or X/Y (male). Each chromosome set has one inherited from the mother and the other from the father.

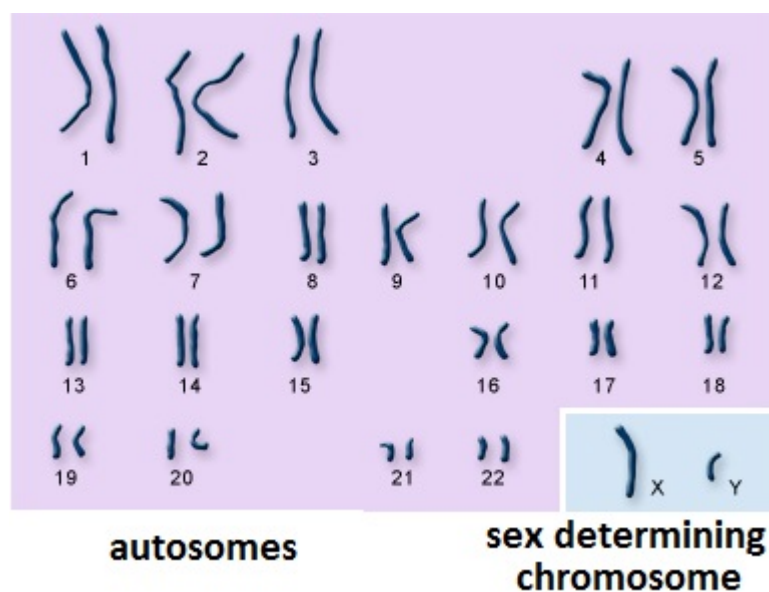


Figure 1-1. Human Genome

(http://www.edinformatics.com/math_science/human_genome.htm).

1.1.2. Deoxyribonucleic Acid (DNA)

Human genome is encoded by deoxyribonucleic acid (DNA). DNA is a double stranded biopolymer composed of nucleotides. Each nucleotide includes a ribose sugar, phosphate group and one of four nitrogenous bases: Thymine (T), Cytosine (C), Guanine (G) or Adenine (A) (Levene, 1919). The nucleotides are paired up to one another (noted as base pair) in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in a backbone. According to base pairing rules, A will only bind with T and C will only bind with G.

1.1.3. Genes

The term *gene* together with the laws of inheritance was illustrated by Gorter and Mendel around 150 years ago (MENDEL, 1950). His famous experiment on pea plant hybridization illustrated that a certain genomic region controlled the inheritabilities of some phenotypic characteristics, such as seed shape and pod colour. A simple definition of gene is a region of DNA that encodes for a protein. The pathway is DNA transcript to RNA and ribonucleic acids (RNA) translate to amino acid. The Encyclopaedia of Genetic Elements (ENCODE) consortium have recently updated definition of a gene as “*a union of genomic sequences encoding a coherent set of potentially overlapping functional products*” (Gerstein et al., 2007). This definition takes into account the considerable genomic complexity in the

human genome, for example, the non-genic regions of the genome which shows high conservation and the enrichment of non-coding RNA.

1.1.4. Gene expression

Gene expression is a procedure in which DNA and genes are used to synthesise gene products, i.e. ribonucleic acid (RNA) and proteins. The Central Dogma of Molecular Biology described this process as a one-way system consists of two steps: firstly DNA is transcribed into RNA, and secondly RNA is translated into proteins (Crick, 1970).

Transcription

Transcription is the first step in gene expression. In this step, the linear nucleotide sequence of DNA is copied by a RNA polymerase into a linear nucleotide sequence in RNA (Strachan and Read, 2011). RNA is classified into several categories based on functions: if the gene is translated (as explained below) into a protein, the gene product is called messenger RNA (mRNA); otherwise, the gene product is called non-coding RNA (ncRNA), i.e. transfer RNA (tRNA) or ribosomal RNA (rRNA) (Storz, 2002).

Translation

After DNA has been transcribed into mRNA, the linear nucleotide sequence of mRNA will be decoded by cellular ribosome three bases at a time. This process will produce a linear sequence of amino acids, and then amino acids will bind into a

protein. Recently, with the understanding of open reading frames (ORF), the new estimate of the number of protein-coding genes are ~ 21,000 (Clamp et al., 2007).

Alternative splicing

The underlying assumption for the above process is called the one gene-one enzyme hypothesis, which assumes that a single gene encodes for one RNA or one protein. In fact, it is possible for a single gene to produce multiple functional transcripts, which is called alternative splicing. There are several different ways where alternative splicing forms, i.e. alternative 5-prime donor, 3-prime acceptor sites and exon skipping (Black, 2003). Chen et al estimated that ~ 90% of genes within human genome go through alternative splicing (Chen et al., 2012).

1.1.5. Genetic Variation

The sequence of human genome is changeable during generations (Strachan and Read, 2011). The change of sequence, called genetic variation, are normally caused by a number of sources: i.e. errors during DNA replication or environmental exposures. Genetic variation provides the basis of natural selection mutations (Duret, 2009). Mutations can either neutral, beneficial or deleterious to people / causal to diseases (Chapter 1.3.1).

The human genome is highly analogous: 99.9% of the entire genome of two unrelated individuals are the same at the nucleotide level (Collins and Mansoura, 2001). The remaining 0.01% genome was classified into three main categories:

single nucleotide polymorphisms (SNPs), Insertions and deletions (termed INDELs) and structural variation.

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are genomic variants changing only one single nucleotide base. Approximately 90% of sequence variants in human are SNPs (Collins et al., 1998). In human genome, SNPs occur approximately once every 1000 bases averagely (Wang et al., 1998). In fact, SNPs are tend to be clustered in the genome instead of randomly distributed (Koboldt et al., 2006).

SNPs fall either within or outside protein-coding regions in the human genome.

There are three possible effects on the translation of a protein sequence of SNPs within protein-coding regions:

1. Silent or synonymous SNPs are variants which change the nucleotide sequence, but do not modify the amino acid sequence of a protein.
2. Missense or non-synonymous SNPs are variants which modify the nucleotide sequence and the amino acid sequence.
3. Nonsense SNPs are variants which change the nucleotide sequence so that the codon is substituted by a codon encoding for a premature stop. The consequence of a nonsense SNP is usually producing of a shorter and unfinished protein or longer, completely destroys function

Although coding variants are important, recent estimates of human exome only accounts for little more than 1% of DNA in the human genome (Maher, 2012). On

the other hand, SNPs falling outside exon regions are called non-coding SNPs. SNPs occur in non-coding regions more frequently than in coding regions, in general, where natural selection is acting and fixating the allele of the SNP that constitutes the most favourable genetic adaptation (Barreiro et al., 2008). Non-coding SNPs have no effect on the translation of a protein sequence. However, the activity and expression of protein-coding genes can be regulated by a variety of DNA elements called regulome, i.e. promoter, enhancer, transcriptional regulatory sequences, regions of chromatin structure and histone modification (ENCODE Project Consortium et al., 2012). It is thought that changes in the regulation of gene activity can disrupt protein production and cell processes and result in disease (ENCODE Project Consortium et al., 2012). Mapping the position of these regulatory elements and how they influence gene transcription could show associations between variations in the expression of certain genes and the development of diseases (ENCODE Project Consortium, 2004).

As described above, the primary goal of the ENCODE project is to determine the role of the remaining component of the genome, much of which was traditionally regarded as "junk" (Maher, 2012).

Insertions and deletions

Insertions and deletions (also note as INDELs) are the second most common class of genetic variation in human genome (Montgomery et al., 2013). They are defined as

a gain/loss of up to 50 nucleotide bases. Similar to SNPs, INDELS can fall within or outside of protein-coding regions of the human genome as well.

INDELS that fall within protein-coding regions have two possible effects on the translation of a protein sequence:

1. INDELS consisting of a number of nucleotides that cannot be evenly divided by three will cause a shift in Open Reading Frame (ORF), which is called frame-shifting (FS). FS will result in translation of a different amino acid sequence.
2. INDELS consisting of a number of nucleotides that can be evenly divided by three will not cause a shift in ORF, which is called non-frame-shifting (NFS). NFS will not alter the translation of the remaining protein sequence.

In addition, INDELS outside protein-coding regions, i.e. intronic and intergenic regions, are called non-coding INDELS. They will not cause any downstream effect on the translation of a protein sequence. But, same as non-coding SNPs, non-coding INDELS can affect the regulation and splicing, which could be associated with the development of diseases.

Structural variation

Structural variations are large variation with in the human genome. Both copy number variants (CNVs), inversions and translocations are in this category (Weischenfeldt et al., 2013).

1. CNVs occur when a segment of DNA sequence appears in variable number of copies compared to the reference genome.
2. Inversions occur when a segment of DNA sequence is reversed in direction in comparison to the rest of the chromosome.
3. Translocations occur when a segment of DNA has modified its position within the genome

Based on their feature, large structural variations are often cause serious phenotypic consequences, which include human diseases such as Down's syndrome-where an extra copy of chromosome 21 is appeared; and DiGeorge syndrome-where a part of chromosome 22 is deleted.

1.1.6. Genetic disorder

One key research field in human genetics is to understand the inheritance of disorders as it occurs in human beings. A genetic disorder is a disease caused by abnormalities of the human genome. It can be classified into two classes based on the number of gene involved in the disorder.

Monogenic disorder

A monogenic disorder is the result of a single mutated gene. In reality, monogenic disorders are often linked to serious phenotypes, so they are tend to be rare in the population.

They normally show familial aggregation with Mendelian pattern (term *Mendelian disease*) of inheritance model, which include:

1. Autosomal dominant, where only one mutated copy of the gene will be needed for a person to be affected by an autosomal disorder. Each affected patient usually has one affected parent. The chance that a child inherits the mutated gene is 50%. Autosomal dominant disorders sometimes have reduced penetrance, which means the percentage of individuals who inherit the mutation develop the disease is less than 100%, i.e. Huntington's disease (Kay et al., 2014).
2. Autosomal recessive, where two mutated copies of the gene are needed for a person to be affected by an autosomal disorder. In the family pedigree, an affected child usually has unaffected parents who carry a single copy of the mutated gene (called carriers). For a child of two carriers, the chance of him/her affected by the disorder is 25%. Example of autosomal recessive disorder includes wet versus dry earwax (Yoshiura et al., 2006) (Rodriguez et al., 2013).
3. X-linked disorders, which are caused by variation on the X chromosome. Males and females are both influenced by these disorders. But males will be affected by them more seriously since males have only one copy of the X chromosome. Examples of X-linked dominant disorders include Rett syndrome (Chattopadhyay and Arora, 2014) and Aicardi syndrome (Prontera et al., 2013). For X-linked recessive disorders, the sons of a man

with an X-linked recessive disorder will not inherit the disorder, but his daughters will inherit one copy of the variant. For a woman carrier, the chance of her son/daughter will carry the copy of the variant is 50%. Example of X-linked recessive disorder includes some serious diseases, i.e. hemophilia A (Aledort et al., 2014)

4. Y-linked disorders, where are caused by mutations on the Y chromosome. This disorder can only be passed from the men to their sons. Y-linked disorders are rare. A typically Y-linked disorder is infertility (Krausz and Chianese, 2014)

Individual monogenic diseases are relatively rare, and affect fewer than 200,000 people for each disorder. However, the most famous database for Mendelian disorders-Online Mendelian Inheritance in Man (OMIM) have involved 5188 phenotypes, of which the molecular basis is known until early 2014 (<http://www.ncbi.nlm.nih.gov/omim>).

Polygenic inheritance

The majority of human diseases and traits with high prevalence are polygenic inheritance (term multifactorial or complex inheritance), i.e. cardiovascular disease, cancer and so on. They can be caused by the effects of multiple genes together with the effect of environmental and behavioural factors. Unlike monogenic disorder, researchers cannot classify their pattern of inheritance simply into several

categories. Polygenic inheritance patterns normally follow a normal (bell-shaped) distribution curve. It shows continuous variation

In early 1900s, Fisher claimed that the heritability of quantitative human traits can be explained by Mendelian patterns of inheritance if they were the result of a combination of variation at many genetic loci and the interaction between loci, including the influence of environmental factors which affect these traits (Fisher, 1918) (Plomin et al., 2009).

An classic example of polygenic inheritance is the human skin colour. Globally, we can observe continuous variation in skin colours. The colour of the skin is the result of pigments, such as melanin. As far as we know, the skin colour is controlled by at least three separate genes, each with their own alleles. Assume that three "dominant" capital letter genes (A, B and C) control dark pigmentation because more melanin is produced. The "recessive" alleles of these three genes (a, b & c) control light pigmentation because lower amounts of melanin are produced. The words dominant and recessive are placed in quotation marks because these pairs of alleles are not truly dominant and recessive as in some of the garden pea traits that Gregor Mendel studied. A genotype with all "dominant" capital genes (AABBCC) has the maximum amount of melanin and very dark skin. A genotype with all "recessive" small case genes (aabbcc) has the lowest amount of melanin and very light skin. Each "dominant" capital gene produces one unit of color, so that a wide range of intermediate skin colors are produced, depending on the number of "dominant" capital genes in the genotype. For example, a genotype with three

"dominant" capital genes and three small case "recessive" genes (AaBbCc) has a medium amount of melanin and an intermediate skin color. This latter genotype would be characteristic of a mulatto. In the following cross between two mulatto genotypes (AaBbCc x AaBbCc), each parent produces eight different types of haplotypes and these haplotypes combine with each other in 64 different ways resulting in a total of seven skin colours. The skin colours can be represented by the number of capital letters, ranging from zero (no capital letters) to six (all capital letters).

1.2. Population genetics concepts

Population genetics is the study of the distributions and changes of allele frequency in a population. Changes of allele frequency can be either caused by chromosomal recombination during meiosis or by mutations (Hartl, 2007).

1.2.1. Allele frequencies

Allele frequency, which is the core concept in population genetics, represents the proportion of a particular alleles (of a genetic variant) at a given locus in the allele pool of a certain population (Klug, 2012).

Hardy-Weinberg Equilibrium (HWE)

Hardy-Weinberg Equilibrium is a common assumption for population genetics, which states that frequencies of allele and genotype-which is the genetic makeup of a

person-in a population will remain constant from one generation to next generation given that there is no other evolutionary influence. Other influences include non-random mating, mutation, selection, genetic drift, gene flow and meiotic drive. But in fact, one or more of these influences are typically present in real populations (Ewens, 2004). In a simple example of one single locus with two alleles, A and a. Assume the allele frequency of allele A is:

$$f(A) = p$$

And allele frequency of allele a is:

$$f(a) = q$$

Under HWE, the estimated genotype frequencies then become:

$$f(AA) = p^2 \text{ for the AA homozygotes,}$$

$$f(aa) = q^2 \text{ for the aa homozygotes,}$$

$$f(Aa) = 2pq \text{ or } f(aA) = 2pq \text{ for the heterozygotes.}$$

The sum of three genotypes is:

$$p^2 + 2pq + q^2 = (p + q)^2$$

Since the sum of the frequencies of allele A and a are equal to 1:

$$p + q = 1$$

Then the sum of the genotypes will be equal to 1 as well. These genotype proportions (p^2 , $2pq$, and q^2) are called the Hardy-Weinberg proportions.

Calculation of allele frequencies

For biallelic SNPs (where the locus only has two alleles), the allele frequencies of the locus can be calculated as a reverse calculation of HWE genotype proportions:

If $f(AA)$, $f(Aa)$ and $f(aa)$ are the three genotype frequencies at a locus with two alleles, then the frequency p of the A-allele and the frequency q of the a-allele are obtained by calculating alleles counts. Because each homozygote AA consists only of A-alleles, and because half of the alleles of each heterozygote Aa are A-alleles, the total frequency p of A-alleles in the population is calculated as follows:

$$p = f(AA) + \frac{1}{2}f(Aa) = \text{frequency of } A$$

In the same way, the frequency q of the a-allele can be calculated as:

$$q = f(aa) + \frac{1}{2}f(Aa) = \text{frequency of } a$$

As explained above p and q sum to 1, so:

$$q = 1 - p$$

$$p = 1 - q$$

Based on allele frequencies, genetic variants can be classified into several categories: variants with minor allele frequencies (MAF) $> 1\%$ are called common variants; variants with MAF $< 1\%$ are normally called rare variants. More specifically, Alharbi et al have defined the rarer variants with MAF between 0.1%- 5% as “paucimorphism” (Alharbi et al., 2005).

1.2.2. Disease-variant Hypothesis and missing heritability

Thousands of genome-wide association (GWA) studies (details in Chapter 1.3.4) have been carried out with the common disease common variant (CDCV) hypothesis. It assumes that common genetic variants, each with small effect (Reich and Lander, 2001), sum up to account for a reasonable proportion of the variance in the phenotype (Lander, 1996). In addition, as discussed above the selective disadvantage conferred by highly penetrant monogenic mutations results in their frequency within the population remaining low (Figure 1-2).

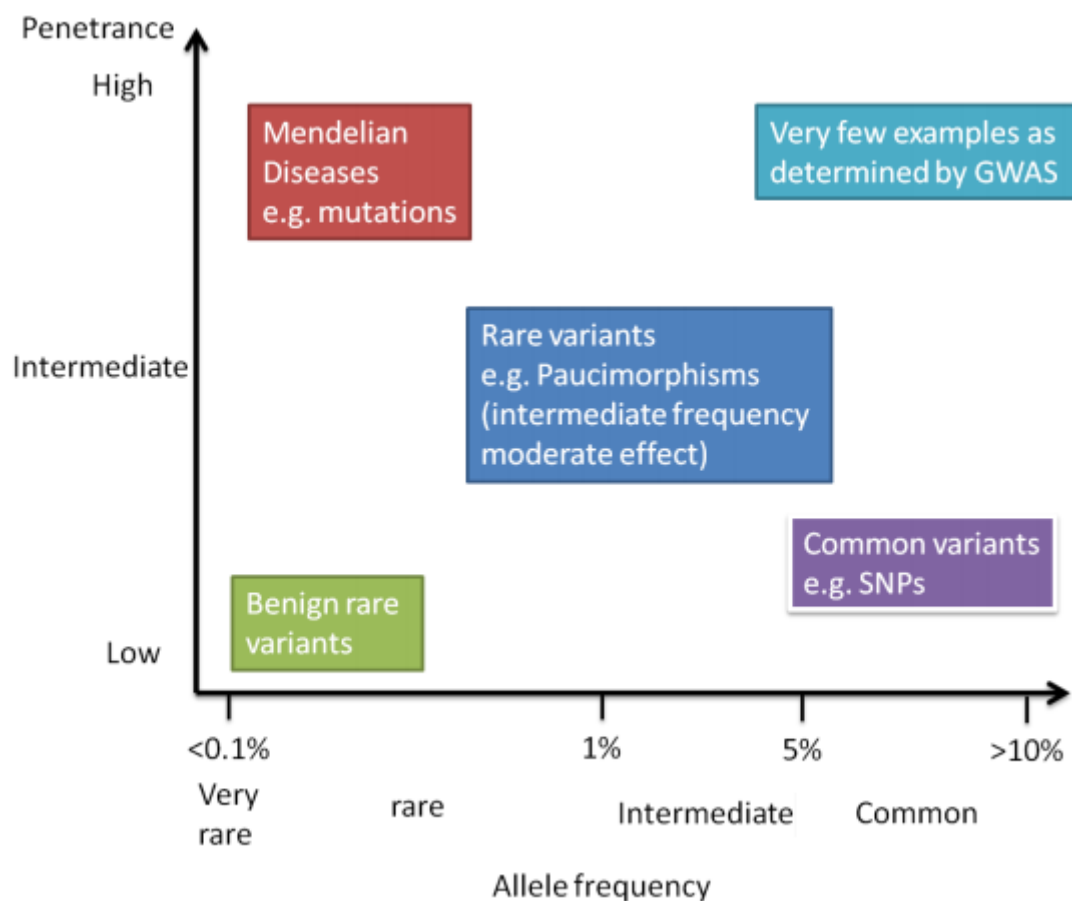


Figure 1-2. Relationship between allele frequencies of genetic variants and penetrance of diseases/traits.

Typically, only variants with minor allele frequency (MAF) greater than 1% are followed up in such studies. There are two reasons for this cut-off:

Firstly, there is an inverse relationship between sample size and the MAF that maximizes the power to detect a true association (Figure 1-3) (Gorlov et al., 2008).

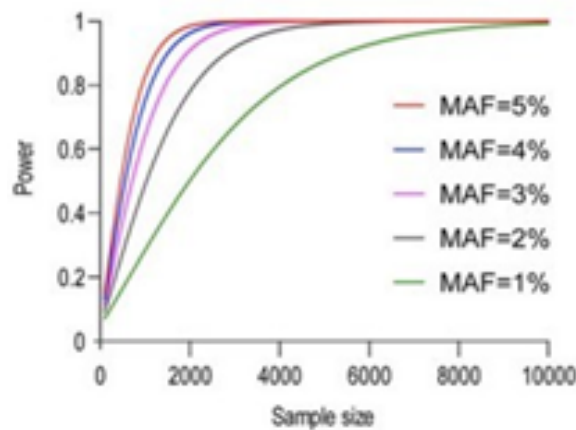


Figure 1-3. Relationship between power and sample size with different MAF (Gorlov et al., 2008)

Secondly, as listed in Table 1-1, SNP genotyping panels are typically designed for common SNPs, therefore they contain a relatively small number of rare variants (Asimit and Zeggini, 2010). These common variants explain at most 5%–10% of the heritable component of disease (Maher, 2008).

	Affymetrix	Affymetrix	Illumina	Illumina	Illumina	Illumina
Platform	500k	6.0M	370k	550k	610k	1.2M
MAF<0.05	55k	106k	9k	32k	35k	62k
MAF<0.01	17k	35k	1k	7k	8k	22k

Table 1-1. Approximate-low frequency / rare variant GWAS platform content. K

means thousand and M means million (Asimit and Zeggini, 2010).

To address the “missing heritability” problem, researchers further developed Fisher’s polygenic inheritance theory (Chapter 1.1.6), which assume that many small genetic effects (most of them not reach the genome-wide significance) working together influence the total heritability of complex traits in human populations since 2010. Based on this theory, Peter Visscher’s lab developed a user friendly method called Genome-wide Complex Trait Analysis (GCTA), which estimate the proportion of variance for complex traits using a restricted maximum likelihood model (REML) (Yang et al, 2011). The same group estimated the proportion of variance for human height explained by SNPs across the whole genome on 3,925 unrelated individuals using REML, and validated the estimation method with simulations based on the observed genotype data. The analysis show that 45% of variance can be explained by considering all SNPs simultaneously. They concluded that most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests (Yang et al, 2010). This method together with the example on height open a door for population genetics field. Another major application of the polygenic inheritance theory is polygenic score. The first successful application of polygenic score analysis to GWAS data was in schizophrenia. Later in 2013, Frank Dudbridge estimated the power and predictive accuracy of polygenic risk score (Dudbridge, 2013). Furthermore, Speed et al claimed that if variance explained is correlated with MAF, the estimates of heritability and genetic covariance can be biased and developed a

method, LDAK, to address such issue (Speed et al, 2012). Utilising a similarly idea, Benjamin Neale's group developed an approach, LD Score regression, that quantifies the contribution of each SNP by examining the relationship between test statistics and linkage disequilibrium (LD). The LD Score regression intercept can be used to estimate a more powerful and accurate correction factor than genomic control. The model is a simple fixed effect model, which regress test statistics (square of Z score) from GWAS against LD score (Bulik-Sullivan et al, 2015 a). Another advantage of such setting is that LD score regression can be applied to meta-analysis summary statistics directly. The meta-analysis data is also the major resource I will going to use in my work, which I will explained in more detail in Chapter 1.3.6. The LD score regression model was further developed so that it can process the bivariate genetic correlation analysis later on (Bulik-Sullivan et al, 2015 b).

In contrast to common variants, rare variants have larger effects than common ones. As shown in Figure 1-4, the advantage of testing rare variants association is that the effects of rare variants tend to be larger than those of higher frequency SNPs (Bodmer and Bonilla, 2008). In addition, the identification of rare variants may facilitate pinpointing causality. It will be easier to infer causality at a locus that contains both common and rare disease-associated variants.

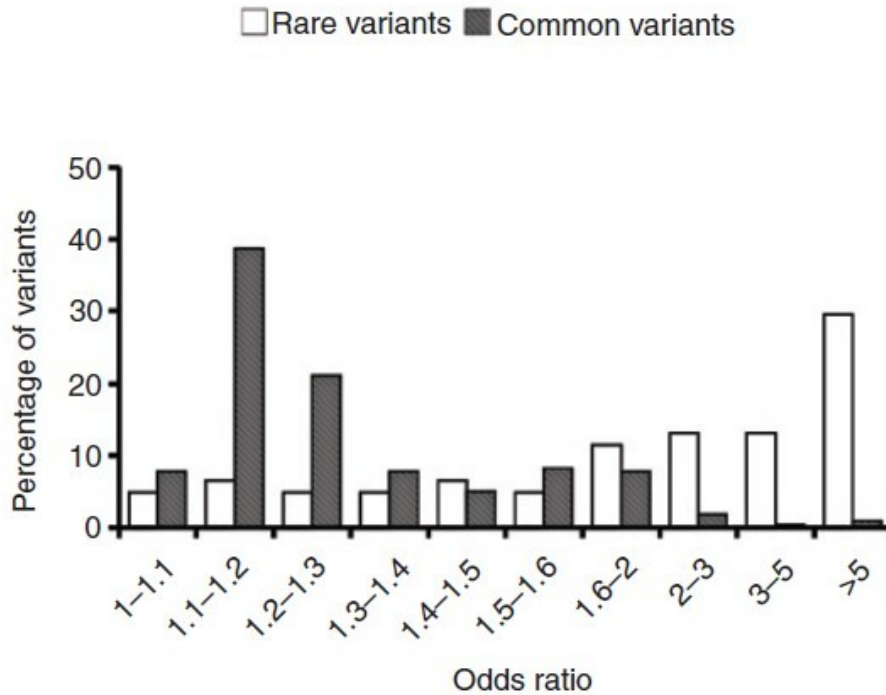


Figure 1-4. Comparison of odds ratio between rare and common variants (Bodmer and Bonilla, 2008).

1.2.3. Linkage disequilibrium, haplotypes and haplotype phasing

A central issue in population genetics is whether it is possible to map causal genes to diseases. Linkage disequilibrium (LD) plays a key role in this issue.

Linkage disequilibrium

LD refers to correlation between neighboring alleles, reflecting combination of alleles (noted as haplotypes) descended from single, ancestral chromosomes (Reich et al., 2001). Several factors will affect the degree of linkage disequilibrium, which including genetic linkage, selection, the rate of recombination, the rate of mutation, genetic drift, non-random mating, and population structure. To

estimate the amount linkage disequilibrium, we used the difference between observed allelic frequencies and the expected allelic frequencies from a randomly distributed population.

If the haplotype frequencies of several genetic variants fit the expected proportions, they are noted as in linkage equilibrium to each other.

Haplotypes and haplotype phase

In population genetics field, a haplotype is a combination of alleles at a certain genomic region (or loci) (Reich et al., 2001). A haplotype can be one locus, several loci, or even an entire chromosome depending on the length of genomic region we are interested in and number of recombination events that have occurred inside the region.

For a simplest model containing two SNPs, SNP A and SNP B, the haplotype will be A1B1, A1B2, A2B1 and A2B2 with haplotype frequencies X11, X12, X21 and X22 respectively. The allele frequencies of A1 A2 B1 and B2 are p1, p2 q1 and q2 respectively (Table 1-2).

	A1	A2	Total
B1	$X_{11}=p_1q_1+D$	$X_{21}=p_2q_1-D$	q1
B2	$X_{12}=p_1q_2-D$	$X_{22}=p_2q_2+D$	q2
Total	p1	p2	1

Table 1-2. Example of a 2X2 haplotype table

If SNP A and SNP B are in LD, X11 is equal to $p_1 * q_1 + D$, where D is the quantity of

linkage disequilibrium between the two SNPs. A more common measure of LD is called D' , which is the normalized D (Lewontin, 1964). D' is represented as:

$$D' = \frac{D}{D_{max}}$$

Where $D_{max} = \begin{cases} \min(p1 * q1, p2 * q2) & \text{when } D < 0 \\ \min(p1 * q2, p2 * q1) & \text{when } D > 0 \end{cases}$

Another commonly used measure of LD is called r^2 (Hill and Robertson, 1968). R^2 is the correlation coefficient between a pair of SNPs, which is equal to:

$$r^2 = \frac{D^2}{p1 * p2 * q1 * q2}$$

For a certain GWAS, if individual-level genotype data is available, the four haplotype frequencies (denoted as X_{11} , X_{12} , X_{21} and X_{22} in this chapter) can be easily calculated using programs such as PLINK (Purcell et al., 2007), CubeX (Gaunt et al., 2007) and MIDAS (Gaunt et al., 2006).

For the same 2 SNPs model shown in Table 1-2, assume locus 1 has alleles A or T and the locus 2 has G or C. For each locus, there are three possible genotypes: AA, AT and TT for locus 1; GG, GC and CC for locus 2. For a certain individual, the genotype may not define the haplotype uniquely, since there are 10 possible genotypes but there are only 9 possible haplotypes at these two loci (Table 1-3). The double heterozygote A_1B_1 - A_2B_2 have two possible genotypes (either AG-TC or AC-TG). This unknown middle cell is called haplotype uncertainty and the technique to infer the haplotype is therefore called haplotype phase.

Locus 1

			A1A1	A1B1	B1B1
			(AA)	(AT)	(TT)
	A2A2	(GG)	AG AG	AG TG	TG TG
Locus 2	A2B2	(GC)	AG AC	AG TC or AC TG	TG TC
	B2B2	(CC)	AC AC	AC TC	TC TC

Table 1-3. *Haplotypes and haplotype uncertainty*

Haplotype phase can be generated through laboratory-based experimental methods, or it can be estimated using computational approaches. The lab-based methods need to spend extra fees to genotype members in the family. On the other hand, computational approaches can phase haplotypes accurately from genotypes without any further cost (haplotype phasing tool will be introduced in Chapter 2).

Haplotypes phasing tools

As described in Chapter 1, haplotype phasing is an important process to estimate the haplotype uncertainty. In this chapter, I shall describe some well-known computational haplotype phasing tools, including PHASE (Stephens et al., 2001), BEAGLE (Browning and Browning, 2007), MACH (Li et al., 2010) and IMPUTE2 (Howie et al., 2009).

PHASE

PHASE is a haplotype phasing algorithm, which is useful for small genomic regions (up to couple hundred SNPs) and small sample sizes (up to one thousand

individuals) but relatively slow compared to the other tools I shall describe below (Stephens et al., 2001) (Stephens and Scheet, 2005). PHASE was considered to be the gold standard amongst population-based haplotype phasing algorithms for a long time (Marchini et al., 2006) before updated tools were developed. Two important updates / refinements of PHASE algorithm is fastPHASE (Scheet and Stephens, 2006) and SHAPE-IT (Delaneau et al., 2012). FastPHASE is the first method that is capable of phasing genome-wide SNP array data. SHAPE-IT includes PHASE algorithm but much faster and can be applied to high-throughput genotyping chips data and whole genome sequencing data.

BEAGLE

BEAGLE is a hidden Markov model (HMM) based algorithm to infer haplotypes (Browning and Browning, 2007). It is faster and more accurate than fastPHASE for medium and large sample sizes (i.e. more than 1000 individuals) but it is not as accurate as fastPHASE for small sample sizes (100 or less individuals). This is because fastPHASE uses Expectation-Maximization (E-M) algorithm which is more accurate than the HMM when a small number of SNPs in a short gene are to be studied. When the number of SNPs in the model increase, the number of haplotype will increase rapidly. This will cause two problems, a) the haplotype frequencies of rare variants will be too small to estimate directly; b) the computational time will become intractable since all haplotypes need be considered (Browning and Browning, 2007).

MACH and IMPUTE2

MACH (Li et al., 2010) and IMPUTE2 (Howie et al., 2009) are designed for genotype imputation of untyped variants (I will introduce imputation in the next section).

However, these algorithms can also be used for haplotype phasing. These two methods can be used for much larger datasets than PHASE and can return more accurate results for large sample sizes when compared to fastPHASE. The accuracy of both methods can be improved by increasing the complexity of the model, for example increase the number of HMM states so that the methods can make better use of the information in the data. But that will cost more computational time (Browning and Browning, 2011).

1.2.4. Genotype Imputation

The success of HapMap and 1000 Genome project has enabled researches to identify tagging SNPs (smaller subset of SNPs) across the whole human genome.

Using HapMap and 1000 Genome data as reference panel, we can now then impute the haplotypes, haplotype frequencies and allele frequencies of hundreds of thousands un-typed SNPs. This process is called genotype imputation. As, shown in Figure 1-5, for unrelated individuals in a population, the raw data consist of a set of genotyped SNPs that has a large number of SNPs without any genotype data (part a). Testing for association at just these SNPs may not lead to a significant association (part b). Imputation attempts to predict these missing genotypes.

Algorithms differ in their details but all essentially involve phasing each individual in the study at the typed SNPs. The figure highlights three phased individuals (part c). These haplotypes are compared to the dense haplotypes in the reference panel (part d). Strand alignment between data sets must be done before this comparison takes place. The phased study haplotypes have been coloured according to which reference haplotypes they match. This highlights the idea implicit in most phasing and imputation models that the haplotypes of a given individual are modelled as a mosaic of haplotypes of other individuals. Missing genotypes in the study sample are then imputed using those matching haplotypes in the reference set (part e). In real examples, the genotypes are imputed with uncertainty and a probability distribution over all three possible genotypes is produced. It is necessary to take account of this uncertainty in any downstream analysis of the imputed data. Testing these imputed SNPs can lead to more significant associations (part f) and a more detailed view of associated regions.

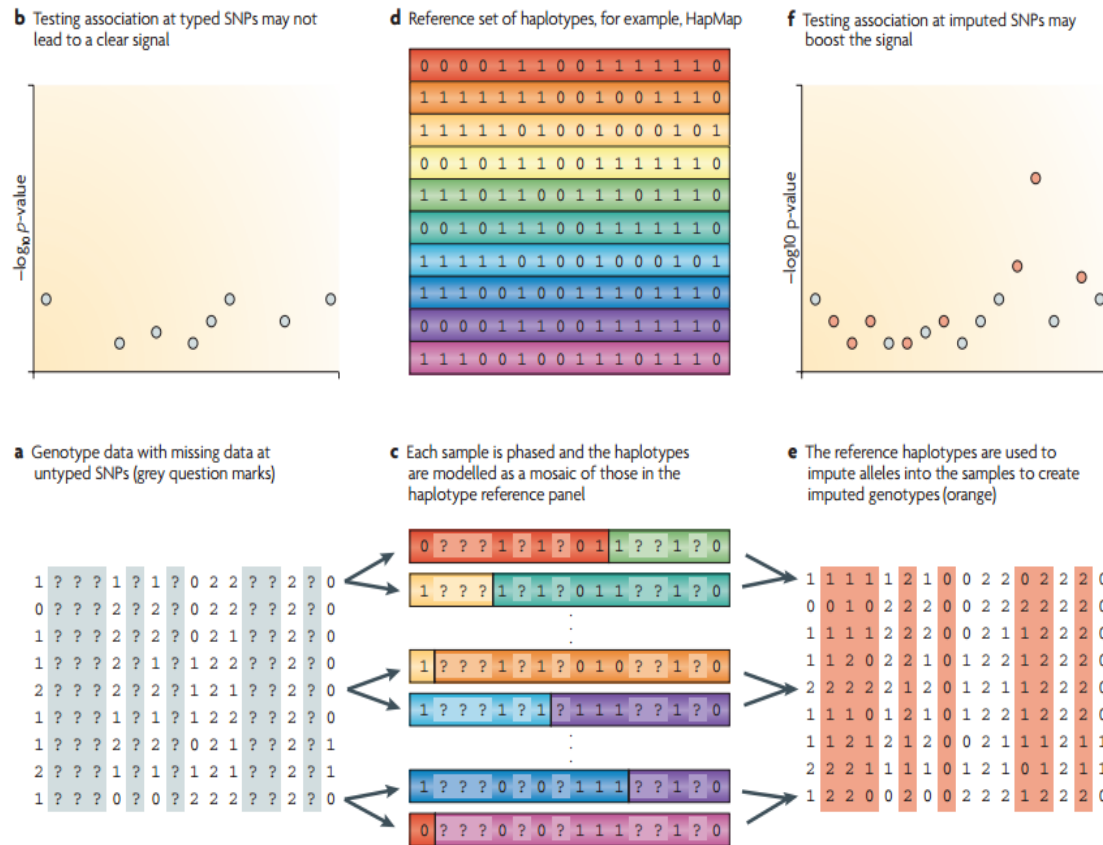


Figure 1-5. Example of genotype imputation (Marchini and Howie, 2010)

Imputation has now become an essential process during GWAS studies. Imputation has the potential to increase the power of GWAS and provide a high resolution overview of an association signals across a locus (which is meaningful for fine mapping) and is particularly useful for combining the GWAS results across studies that rely on different genotyping platforms (i.e. meta-analysis) (Li et al., 2009).

We can classify imputation tools into two categories: computationally intensive tools and computationally efficient tools.

Computationally intensive tools: MACH, minimac and IMPUTE2

Computational intensive tools, such as MACH (Li et al., 2010) and IMPUTE2 (Howie et al., 2009), take into account all observed genotypes when imputing each missing genotype.

Typically, tools that consider all available markers and all available haplotypes require intensive computation but do better at estimating missing genotypes, particularly for rare SNPs (Li et al., 2009). Both MACH and IMPUTE2 are based on coalescent models to impute the missing genotypes. These models are based on the fact that new haplotypes are developed from old haplotypes by the processes of mutation and recombination. Since mutations and recombination are quite rare inside small genomic regions, haplotypes tend to look similar to each other. So, for instance, if we see two haplotypes, e.g. 1000 and 0010 in a sample (where 0 and 1 are the two possible alleles at the four loci), we have better chance of observing the haplotype 1010 (caused by recombination) or 1100 (caused by mutation) than observing the haplotype 1110 (caused by both a recombination and a mutation) (Browning and Browning, 2011). The approximate coalescent gives rise to a HMM, and its parameters are estimated with the use of iterative algorithms such as the stochastic E-M algorithm (Tregouet et al., 2004). One advantage of IMPUTE2 is that it can include both HapMap and 1000 Genome as reference panel (Howie et al., 2009).

Minimac (Howie et al., 2012) is a low memory, computationally efficient implementation of the MACH algorithm. Now, a two-step procedure is recommended. The first step is a pre-phasing step using MACH. This step is a time-consuming step but it is only needed once. The second step is the actual imputation into phased haplotypes using minimac and is very fast (i.e. one million markers for 1000 individuals using 100 reference haplotypes takes about an hour, and computing time increases linearly with all these parameters) (http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook).

Computationally efficient tools: PLINK and BEAGLE

Computationally efficient tools, like PLINK (Purcell et al., 2007) and BEAGLE (Browning and Browning, 2007), typically only focus on genotypes for a small number of nearby markers when imputing each missing genotype. PLINK used an E-M algorithm for genotype imputation function while BEAGLE used an HMM model to impute the missing genotype.

Performance comparison

In a comprehensive comparison between BEAGLE, IMPUTE, MACH and PLINK, Nothnagel M et al concluded that BEAGLE, IMPUTE, and MACH yielded nearly identical trade-offs between imputation accuracy and efficacy whereas PLINK performed consistently poorer (Nothnagel et al., 2009). A more recent comparison made by “GoldenHelix” compared BEAGLE, IMPUTE2 and Minimac for accuracy,

computational time and memory usage (<http://blog.goldenhelix.com/?p=1911>).

They concluded that IMPUTE2 is the most accurate method; however one of the disadvantages of the method was it's the intensive memory usage. BEAGLE and Minimac use much less memory but need more computational time. Besides, pre-phasing will decrease the accuracy while save a lot of time. On the other hand, BEAGLE seems like the most user friendly program with good manual. So choosing the most appropriate imputation programme is a trade-off between imputation quality, computational time and memory usage.

1.3. Mapping genetic variants to complex traits

For each human being, on average approximately 3.5 million SNPs and 610,000 INDELs have been identified in the whole genome averagely (Pelak et al., 2010).

Characterising the relationship between genetic variants and complex human disease/trait susceptibilities will lead to a better understanding of the aetiologies of human diseases. Here I will explain the background knowledge of genetic mapping first.

1.3.1. Phenotypes and phenotype standardization

In human genetics field, a phenotype is a composite of a human being's observable characteristics or traits, i.e. type 2 diabetes or height.

A phenotype is normally caused by the expression of human genes, the influence of environmental factors, and their interactions:

Genotype (G) + environment (E) + genotype & environment interactions →
phenotype (P)

Almost all of the traits can fall into two major distinct classes: binary outcomes and continuous traits.

Binary outcomes (disease status)

Binary data is a data type which can take on only two possible values: i.e. "success" or "failure", "yes" or "no", "true" or "false" and "present" or "absent". In the human genetics field, binary outcomes often represent the disease status.

Numerically, these two values are normally coded as 0 and 1, where 0 means absent and 1 means present. Binary data is the simplest type of categorical data.

Randomly selected binary variables follow the Bernoulli distribution. Regression analysis on/given outcomes that are binary variables is mainly accomplished through logistic regression.

Continuous traits (quantitative traits)

Most of traits do not fall into discrete classes. Rather, when a certain population is analysed, a continuous distribution of phenotypes can be identified. These continuous values are called quantitative traits. Many human phenotypes can be classified into this class, which including height, weight, blood pressure and so on. Regression analysis on predicted traits that are continuous variables is majorly accomplished through linear regression. Most quantitative traits can be attributed to polygenic disorders, i.e. product of multiple genes and environmental factors. Quantitative traits from random selection need follows the approximately normal distribution so that linear regression can be used to analyse these traits.

Standardization

In probability theory, the normal (or Gaussian) distribution is a very common continuous probability distribution. The normal distribution is sometimes informally called the bell curve. For most of the statistical analysis in population genetics field, we need a phenotype which follows an approximate normal distribution. For raw phenotype data which are not normal distributed, we can standardise them using the following ways:

1. Standard score, also called Z-score, is the number of standard deviations an observation is away from the mean. The use of “Z” is because the normal distribution is also called the “Z distribution”.

The Z-score of a raw value x is

$$Z = \frac{X - \mu}{\sigma}$$

Where μ is the mean of the population and σ is the standard deviation of the population.

The value of z represents the difference between the observed value and the mean of the population in units of the standard deviation. Z is negative when the raw value is smaller than the mean and positive when bigger than the mean.

One advantage of Z-score is that it does not have a unit, therefore we can compare traits with different units after we standardized them.

2. Logarithm transformation, the logarithm of a number is the exponent to which the base must be raised to produce that number.

$$y = b^x \Leftrightarrow \log_b(y)$$

The logarithm transformation can only be applied when data are strictly positive. Two kinds of logarithms are used frequently: “natural” logarithms (irrational number e as its base) and base-10 logarithms.

The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making the data more interpretable and for helping to meet the assumptions of inferential statistics.

1.3.2. Association analysis for candidate genes

Another classic method, which is still a useful follow-up approach nowadays, is the candidate gene analysis (Westaway et al., 2011). The scientists will select a gene (or a group of candidate genes) based on their biological knowledge and prior experience. Then they will sequence the gene (genes) for genetic variants in a group of people with / without a phenotype (or some relative phenotypes). Finally, they will try to identify the association between the variants and phenotypes. This approach is highly possible to be failed in certain genes and also lacks of replication since the knowledge on the molecular mechanisms of these phenotype is uncertain (Richardson et al., 2013).

1.3.3. Genome-wide association study (GWAS)

Genome-wide association study (GWAS) extend the association analysis approach to the whole human genome. The formal definition of GWAS is a statistical test which are performed using dense maps of common SNPs (or note as single nucleotide polymorphism-SNP) on a group of participants from a study. The study can be either a disease-based cases-control design or exposure-based cohort design.

GWAS offered the opportunity for the “hypothesis-free” analysis to understand the disease pathophysiology, which will overcome the difficulties and obstacles of the

traditional candidate gene approach (Kitsios and Zintzaras, 2009). Since GWAS includes a huge amount of genetic markers, it is more likely to identify the truly causal variant or the proxy variants in strong LD with the causal variants. But the side effect of GWAS-multiple testing-will increase the chance of false positive findings. In statistics, the multiple comparisons, multiplicity or multiple testing problem occurs when one considers a set of statistical inferences simultaneously (Benjamini, 2010). Bonferroni correction is the most common way to counteract the problem of multiple testing (Bonferroni, 1936), which is simply divide the test-wise significance level (α) by the number of tests (N), the significance level after correction will be:

$$\alpha_{\beta} = \frac{\alpha}{N}$$

For GWAS, N is the number of SNPs involved in the test. So given the large multiple testing in GWAS, very few signals exceed the genome-wide significance threshold and those that do not exceed this stringent statistical requirement are generally neglected (Kitsios and Zintzaras, 2009).

The first GWAS were proposed by Risch and Merikangas in 1996. This study showed that the tests between common genetic variants were better powered than linkage analysis to identify causal variants (Risch and Merikangas, 1996). Later, scientists found that genetic effects underlying complex traits and common human diseases are small, and their detection requires typing of single nucleotide polymorphisms (SNPs) in large samples (Cardon and Bell, 2001) (Colhoun et al., 2003). Then, the completion of the Human Genome Project provided a platform for identifying and

characterizing common genome-wide genetic variation (International Human Genome Sequencing Consortium, 2004). After that, GWASs have identified thousands of single nucleotide polymorphisms (SNPs) associated with human complex traits and diseases (Hindorff et al., 2009) (Manolio, 2010). The National Human Genome Research Institute has established a catalog of published GWAS studies (<https://www.genome.gov/gwastudies>). As of 01/10/14, the catalog includes 1788 publications and 12329 SNPs.

1.3.4. Regression analysis

In this section, I will explain the concepts of regression analysis since association analysis, such as GWAS, is highly related to it.

Linear regression

In statistics, linear regression is an approach for identifying the approximate linear relationship between a dependent variable (Y) and one or more explanatory variables (X or $X_1 \dots X_n$). The dependent variable should be a quantitative variable while the explanatory variable can be either continuous or categorical variable.

In linear regression, data are modelled using a linear model. For example, the relationship between an explanatory variable X (i.e. a SNP) and a dependent variable Y (i.e. a certain phenotype) can be identified by a linear regression model:

$$Y = \beta_0 + X\beta_1 + \varepsilon$$

Where β_0 is the intercept of the regression; β_1 is beta coefficient of the regression. ε is the error variable (which means an unobserved random noise to the linear relation).

Figure 1-6 shows an example of the simplest linear regression model. \bar{y} is the mean of the outcome variable. “ Y_i ” is the predict value of the regression for event i . “ y_i ” is the observed value of event i . $y_i - Y_i$ is the residual about regression (noise). $Y_i - \bar{y}$ is the signal due to regression. The green line is the line of the best fit.

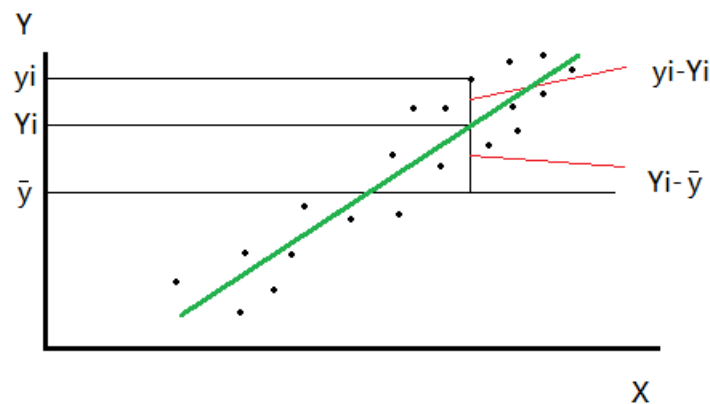


Figure 1-6. Example of scatter plot and linear regression of a continuous explanatory variable. X is the predictor variable (independent variable) and Y is the outcome variable (dependent variable). The regression line is in green.

Logistic regression

On the other hand, logistic regression is a kind of generalized linear model (GLM).

GLM is a flexible generalization of ordinary linear regression that allow for response variable that have distribution other than a normal distribution. Like linear regression, it also predicts the linear relationship between a response variables (Y)

and one or more explanatory variable(s) (X or $X_1...X_n$). The dependent variable of logistic regression should be a categorical variable (i.e. binary disease status). Meanwhile, the explanatory variables can be either categorical or continuous variables.

Logistic regression uses a link function to link response and explanatory variables. In genetic epidemiology field, the simplest link function is:

$$LN(odds\ ratio) = \beta_0 + X\beta_1 + \varepsilon$$

Where LN is the natural logarithm. β_0 is the intercept of the regression; β_1 is beta coefficient of the regression. ε is the error variable. The odds ratio is the odds of cases divided by the odds of controls for a disease, where odds of an outcome D is given by probability of D occur ($P(D)$) divided by probability of D not occur ($P(\text{not } D)$).

Model fit: least squared method

The least squared method is an approach to the approximate solution of over determined systems. "Least squares" means that the overall solution minimizes the sum of the squares of the errors made in the results of every single equation.

Hypothesis test: testing the significance

The next step is to test the significance of the association using a hypothesis test.

For association analysis, we want to test whether there is an association between a SNP (or between a set of SNPs) and a trait, which means whether the slope (effect size $=\beta_1$) is equal to zero.

So the null hypothesis H_0 will be:

$$\beta_1 = 0$$

While the alternative hypothesis H_1 will be:

$$\beta_1 \neq 0$$

Normally the t statistics (details in Chapter 3) will be an intermediate parameter for the hypothesis test (either one-sided test or two-sided test). A large value of t statistic, either negative or positive, would provide evidence against the null hypothesis. Therefore, the decision rules to reject H_0 if the absolute value of t is greater than the critical value of the student's distribution on n-2 degrees of freedom for a specified confidence level (95%, 99% or even more confident).

Assumption of regression analysis

The following are the major assumptions of regression analysis:

1. Lack of measurement error is important.
2. Linearity, which means the mean of the dependent variable is a linear function of the explanatory variables.
3. Equal variance of errors, where the different dependent variables have the same variance in their error, regardless of the predictor variables.
4. Lack of multicollinearity in the explanatory variables. If two or more explanatory variables are in a strong correlation to each other, this will

cause a problem called multicollinearity for standard least squares estimation approaches.

Example of regression analysis: testing the association of a SNP and a trait

For association analysis of a certain SNP A, the linear relationship between its genotypes and a quantitative phenotype can be represented using a regression model. In the example shown in Figure 1-7, I used an additive model, which assume the effect of the phenotype will increase / decrease a certain amount with the increase of per rare allele. The genotypes are coded as 0, 1 and 2 for TT, TA and AA respectively. Each dot represents the effect on the phenotype of an individual with one of the genotypes. The slope of the green line (line of the best fit) is the beta coefficient. The baseline is the intercept of the regression model.

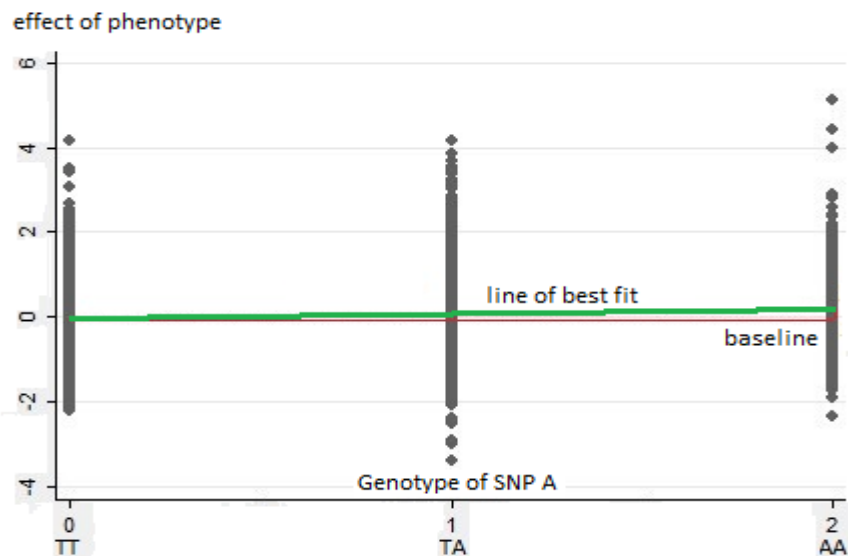


Figure 1-7. Example of association analysis between a biallelic SNP and a quantitative phenotype. TT, TA and AA are the three genotypes of SNP A. The green line is the line of the best fit. The red line is the line of baseline effect.

1.3.5. Meta-analysis of Genome-wide association analysis

One problem with GWAS is that many stand-alone GWAS studies lack sufficient power to obtain precise results (Zeggini and Ioannidis, 2009) (Cambon-Thomsen, 2003) so the synthesis of available open resources and data pooling remains important. In statistics, a meta-analysis refers to methods that compare and combine results from different studies. Meta-analysis is normally done by identification a weighted average of a common measure of effect size (in regression analysis effect size means beta coefficient). The major advantage of meta-analysis is that it increases the sample size so that the precision and accuracy of estimates can be improved. More importantly, the statistical power to detect an effect will increase as well. Commonly used software for meta-analysis include METAL (Willer et al., 2010) and STATA (StataCorp, 2013) and PLINK.

Numerous examples of meta-analyses have identified disease and quantitative trait susceptibility loci that small individual level studies did not identify reliably (Ioannidis et al., 2006). It is, therefore, desirable to utilize the published and on-going studies.

Data synthesis models for meta-analysis

There are several data synthesis approaches for meta-analysis. Two of them are widely for GWAS meta-analysis:

1. Fixed effect model is the most popular approach for synthesizing GWAS data since it is the most powerful approach for discovering phenotype-genotype

association (Pfeiffer et al., 2009). The assumption for fix effect model is that the true effect of each risk allele is the same in each individual study. This assumption is quite hard to achieve, but fixed effect model will maximizing discovery power compared to random effect model (Pereira et al., 2009).

For fixed effect model, the weight of each study is:

$$\text{Weight } (W_i) = \frac{1}{SE_i^2}$$

Where SE_i is the standard error of effect size for each study “i”. Based on this weighting system, bigger studies will get more weight than smaller ones.

Summary effect then will be:

$$\text{Summary effect} = \frac{\sum(W_i * \text{effect})}{\sum W_i}$$

2. Random effect model assumes the true treatment effect varies between studies. Random effect models are not designed for discovering association since they will return far more limited power than fixed effect models. However, random effect models will performance better than fixed effect models when the aim is to synthesize heterogeneous studies and estimate the average effect size of the associated variant and its uncertainty across different populations (Ioannidis et al., 2007).

For random effect model, the weight of each study is:

$$\text{Weight } (W_i) = \frac{1}{(SE^2 + I^2)}$$

Where SE_i is the standard error of effect size for each study “i” and I^2 is the heterogeneity term of the meta-analysis. The definition of heterogeneity is variation between the true intervention effects underlying the different studies.

Heterogeneity can cause by difference between patients, intervention, outcome definition and design of each study of the meta-analysis.

An example of a meta-analysis results is shown in Table 1-4:

ID	A1	A2	Frequencies of A1	b	se	p	N
SNP1	A	G	0.8493	0.0024	0.0055	0.6653	129850
SNP2	C	G	0.0306	0.0034	0.0115	0.7659	129799
SNP3	A	C	0.5128	0.0045	0.0038	0.2319	129830

Table 1-4. *An example of meta-analysis summary statistics.* For each SNP, A1 is the alternative allele and A2 is the risk allele.

Polygenetic inheritance, missing heritability, estimating heritability using REML, LDSC etc

1.3.6. Genetic fine mapping

Current GWAS studies have identified a huge amount of genomic regions that cover causal variants for different complex traits. Although GWAS studies have much greater resolution than linkage analysis, it is still difficult to know the uncertainty between statistical “distance” and physical distance. We need further analysis to refine the location of causal variants as sharply as possible so that follow-up functional studies and hypotheses regarding mechanism can be processed (McCarthy and Hirschhorn, 2008). The process of this kind of further analysis is often noted as “genetic fine mapping.” As shown in Figure 1-8, this terminology is a holdover from family-based linkage analysis.

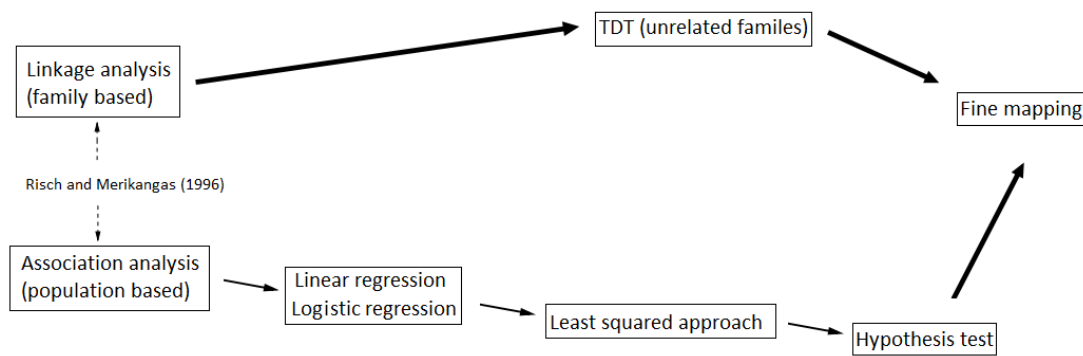


Figure 1-8. Relationship between linkage analysis, association analysis and fine mapping. TDT refers to transmission disequilibrium test, which is a family based association test for unrelated families. It is out of the range of this thesis. For details please read the paper (Spielman et al., 1993).

There is strong LD throughout most of the genome (International HapMap Consortium, 2005), so a genotyped proxy variant and the untyped causal variant could be tens of thousands base pairs away from each other, and the region in between them could contain tons of alleles, none of which are strongly correlated with the causal allele. Moreover, it is quite often that some variants show more or less equivalent evidence of association for a certain association signal. It is also possible that multiple independent association signals exist at a certain locus, for example, protein tyrosine kinase 2b (*PTK2B*) associated with Alzheimer's disease (Lambert et al., 2013). For a set of potential causal variants, fine mapping, which utilises the LD pattern across a region, to test which variant (or subset of variants) is (are) most likely to explain the signal of association.

Fine mapping can not only help narrow down the list of possible causal variants and simplify future functional studies, but also increase the power of association if the

dense set of SNPs use for the GWAS is reasonably high (Wiltshire et al., 2008) (Sanna et al., 2011). In my point of view, common and rare variants that are poorly represented in common genotyping arrays (which means an array with low density but cover the whole genome) may account for a large proportion of trait heritability. Ignoring these variants might exclude identification of casual trait associated loci so that we are underestimating the heritability. Fine mapping (using either a high density gene centric genotyping array or a high coverage sequencing data of candidate regions) should be the logical next step after GWAS.

1.4. Existing model selection approaches for identifying independent SNPs associated with polygenic traits

As explained above, the major aim of genetic fine mapping is to identify the “casual” (or state more clearly the independent) variant(s) in a candidate genomic region. To achieve this aim, we need to select a subset of possible causal variants (term independent SNP) from the entire association analysis results. For such selection procedure, there are four types of statistical approaches exist:

1. Select SNPs for follow up based on their ranked beta coefficient of the single regression analysis (i.e. the most significant SNP, the second most significant SNP and so on)
2. An approximate Bayesian procedure, based on posterior probabilities that each marker is the causal marker.

3. A Least absolute shrinkage and selection operator (Lasso), which is a shrinkage based selection method for linear regression.
4. A stepwise procedure, which recalculates regression coefficient conditional on the most (second most etc.) significant SNP of a genomic region to find out the independent effect SNPs.

The first strategy has drawbacks since the selected markers for follow up are based on statistical significance (rather than the real functional variant). There is high chance that the causal SNP is excluded from the follow up study because of experiments cost, time issue or labor intensive. So I do not recommend using this strategy anymore.

On the other hand, the Bayesian analyses produce an estimate of the probability of association. Bayesian methods can combine information from both GWAS results, SNP information derived from bioinformatics databases, empirical SNP weights, and the researchers' subjective prior opinions. There are different ways to implement Bayesian procedures (Wellcome Trust Case Control Consortium et al., 2012) (Valdar et al., 2012) (Thompson et al., 2013). The advantage of the Bayesian analysis is that Bayes factors, although not as well known as p value, for different SNPs can be compared quantitatively. For instance, for a certain genomic region, if we assume a certain number of causal SNPs in the selected set of SNPs, it is straightforward to calculate the posterior probability that any particular SNP is causal taking into consideration data from the fine-mapping analysis (Wellcome Trust Case Control

Consortium et al., 2012). There are two problems that may prevent a researcher using Bayesian approach: firstly, these approaches are relying on have a reasonable prior distributions which can be difficult to determine (<http://www.bayesian-inference.com/priors>). Bad prior distribution may lead to poor estimation of posterior distribution. Secondly, Bayesian approaches are normally computational expensive and time-consuming (Grzegorzczuk, 2010).

The third strategy a step-wise procedure, is the most commonly used one. Compare to Bayesian analysis, step-wise procedures are computational fast and easy to understand. In the following section, I will introduce some popular step-wise based fine mapping methods used for complex human traits.

The third approach, LASSO, is a relevant new statistical method which is proposed by Tibshirani in 1996 (Tibshirani, 1996). The Lasso minimizes the residual sum of squares subject to the constraint on the sum of absolute value of coefficients. In 2004, Efron et al. proposed the Least Angle Regression(LARS) (Efron et al, 2004) which is a computationally efficient model selection algorithm. There is a close connection between the LARS and the LASSO. A simple modification of the LARS algorithm can yield all the LASSO solutions. Due to their popularity and usefulness, the LASSO and the LARS have drawn intensive research interest in the statistical field. Later in 2005, the Elastic Net proposed by Zou and Hastie (Zou and Hastie, 2005) uses a novel regularization penalty. The naive Elastic Net uses a combination of the LASSO and the Ridge regression penalty. However, the Elastic Net uses a

scaled version of the naive Elastic Net estimate to reduce the over shrinking of parameters. It has been shown that the Elastic Net outperforms the LASSO (Zou and Hastie, 2005). In a GWAS (SNP based association analysis) point of view, The above LASSO methods can be used to identify genetic variants that associated with variation in the phenotypes. Some successful model selection methods using LASSO or using combination of Bayesian approach and LASSO have been developed since then (Wu et al, 2009, Li et al, 2011, Motyer et al, 2011).

1.4.1. Stepwise regression and Conditional analysis

Stepwise regression is a classic scheme for sequentially adding or removing variables from the model. In genetic field, conditional analysis, which is the most commonly used model selection method, has been used as a tool to identify secondary association signals at a locus (Lango Allen et al., 2010) (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011) (Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011).. A more general and comprehensive strategy would be to perform a conditional analysis, starting with the top associated SNP, across the whole genome followed by a stepwise procedure of selecting additional SNPs, one by one, according to their conditional P values.

A more common and expended strategy would be to perform a conditional analysis as follows:

1. Firstly, test the association between a trait and test SNPs conditional on the most significant associated SNP, across a certain genomic region
2. Then, follow a stepwise procedure of selecting additional SNPs, one by one, according to their conditional P values.
3. Finally combine all the reminder SNP in the model and process a step-wise regression to remove SNPs highly correlated to each other.

Such a strategy would allow the discovery of more than two associated SNPs at a locus (Galarneau et al., 2010) (Trynka et al., 2011).

1.4.2. Model selection approaches

Model selection is a way of selecting a good statistical model from a group of candidate models. The word “good” is kind of a balance between fitness and simplicity of the model. Consider the specificity, we need to include the main effects and interactions in the model. On the other hand, parsimonious will reduce the computational time of our analysis.

In order to select the best model, we firstly need a criterion or benchmark to compare the models. We also need a search strategy.

Possible criteria

1. The coefficient of determination (R^2), which measures the goodness of fit of a statistical model. It is not a good criterion since it always increase when extra explanatory variables are added to the model. Adjusted R^2 is an attempt to take account of the phenomenon of the R^2 . It is better than R^2 but it penalized bigger models.
2. Mallow's C_p (Mallows, 1973), which is used to assess the fit of a regression model that has been estimated using ordinary least squares.
3. Akaike's Information Criterion (AIC), named for Akaike Hirotugu, is a measure of the relative quality of a statistical model (Akaike, 1974). It offers a relative estimate of the information lost when a given model is used to represent the whole data.

For any statistical model, the AIC value is:

$$AIC = 2k - 2LN(L)$$

Where k is the number of parameters in the model and L is the maximized value of the likelihood function of the model. LN is the nature logarithm.

4. Bayesian Information Criterion (BIC), which was first developed by Schwarz (Schwarz, 1978), is a criterion for selection among a finite set of models. It is partly based on the likelihood function and is closely related to AIC. In comparison to BIC, AIC penalizes the number of parameters less strongly.

For a large statistical model, the BIC value is:

$$BIC = -2 * LN(L) + k * LN(n)$$

Where n is the sample size. k is the number of free parameters to be estimated.

L is the maximized value of the likelihood function of the model. LN is the nature logarithm.

Search strategies

1. Best subset, where search all possible models and take the model with highest R^2 , lowest C_p or other information criteria.
2. Stepwise (forward, backward or both), where choose an initial model and plus or minus one parameter from the model until the best model is found. It is useful when number of predictors is large.
3. Greedy, where take the biggest jump (up or down) in the selected criterion.

1.5. Materials

In this section I will briefly overview the sample datasets, bioinformatics tools, scripting programme and statistical tools I am going to use in this work.

1.5.1. British Women's Heart and Health Study (BWHHS)

The British Women's Heart and Health Study is a prospective cohort study of cardiovascular disease in older British women (Chapter 9.13 BWHHS). The samples were randomly selected from British women aged 60-79 and were drawn from 23

towns in England, Scotland and Wales. Most of them are European originals.

Information was collected using general practitioner record reviews, a self-completed questionnaire, research nurse interview, and physical examinations.

The study aims to provide information about the existing patterns of treatment for heart disease, and further the understanding of risk factors and disease prevention.

In total, 7173 participants were approached and information pertinent to approximately 60% of participants (4286) is available. 20% of women had a doctor diagnosis of either myocardial infarction, angina, heart failure, stroke, or peripheral vascular disease. 50% of participants were hypertensive and more than 25% of them were obese. 50% have a total cholesterol level greater than 6.5 mmol/l. For all 4286 participants, 3800 have DNA samples (Lawlor et al., 2003). The structure of BWHHS is shown in Figure 1-9.

First phase, medicals and
Interviews (1999/2001)



Published  Research results

Questionnaire (2003)



Published  Research results

Questionnaire (2007)



Published  Research results

Questionnaire and Actigraph (2010)

Figure 1-9. *Structure of the BWHHS study*

BWHHS “Metabohip”

Of the available BWHHS participants, 2024 were genotyped using the Illumina Cardio-Metabohip (called “Metabohip”). The Metabohip is a custom genotyping array design for fine mapping additional loci associated with cardiovascular disease (CVD) related traits. It covers ~200,000 SNPs which are selected based on the knowledge of previous GWAS studies for type 2 diabetes, coronary artery disease,

myocardial infarction, body mass index, glucose and insulin levels, lipid levels and blood pressure (Voight et al., 2012).

BWHHS “Cardiochip”

Of the available BWHHS participants, 3443 were genotyped using 50 k HumanCVD BeadChip (called “Cardiochip”) (Illumina, San Diego, CA). The CardioChip consists of ~50,000 SNP markers that capture genetic regions across ~2,100 genes identified as being associated with CVD from the literature (Keating et al., 2008).

1.5.2. HapMap and the 1000 Genome Project

HapMap

The aim of the international HapMap project is to determine the common patterns of DNA sequence variation within the human genome (a haplotype map, so called “HapMap”) and to make these precious data freely available in the public domain. Using the haplotype map in a genomic region, scientists can then identify tagging SNPs that could be used as representatives to represent haplotypes uniquely (<http://hapmap.ncbi.nlm.nih.gov>) (Chapter 9.13 HapMap). Moreover, HapMap provides a landscape of precious linkage disequilibrium (LD) information of human, which is the foundation for fine-scale analysis of genomic variation. It enhances our ability to fine map associations between common variants and complex diseases / traits (Slatkin, 2008).

The original HapMap project genotyped more than one million of common genetic variants (SNPs) in 270 individuals from four human populations (Chinese, Japanese, European and Nigerian) (International HapMap Consortium, 2003). The allele frequencies and the degree of association (LD) between these SNPs were also measured for each population. At the time of writing, HapMap (Phase 3) has identified ~ 1.6 million SNPs in 1,184 individuals from 11 different populations (International HapMap 3 Consortium et al., 2010). In Chapter 3, I will use the genotype data of HapMap Utah Residents with Northern and Western European ancestry (CEU) and Toscana in Italia (TSI) population to process an admixture population simulation.

1000 Genomes

The 1000 Genome project has similar goals to the HapMap project, which is to provide a well-annotated haplotype reference map (1000 Genomes Project Consortium et al., 2010) (Chapter 9.13 1000 Genomes). In contrast to the HapMap project, the 1000 Genomes project has sequenced more than 1000 individuals from 26 different populations in order to enable them to discover low frequencies variants (MAF between 1% and 5%). The 26 populations used within the 1000 Genomes project are from many different locations around the world. They can be classified into 5 super populations: African, admixed American, East Asian, European and South Asian. In order to reduce costs, the 1000 Genomes project opted to use low-coverage whole genome sequencing (WGS) (~5X). Although variants with low frequencies can be identified using low-coverage WGS, but with

such limited sample size and coverage issues, the approach is limited to detecting variants with a population allele frequency of ~1% (1000 Genomes Project Consortium et al., 2010). In this work, genotype and haplotype data of 1000 Genome CEU and British in England and Scotland (GBR) population will be used as reference genotype data in Chapter 5 for a performance comparison of fine mapping methods. I will also use them as a reference panel to impute genotypes of BWHHS individuals in Chapter 6.

1.5.3. Avon Longitudinal Study of Parents and Children (ALSPAC)

The Avon Longitudinal Study of Parents and Children (ALSPAC), which is also known as Children of the 90s, is a population-based cohort study (Fraser et al., 2013) (Chapter 9.13 ALSPAC). The major aim of the study is to establish a trans-generational, prospective, observational birth cohort and use it to understand how genetic and environmental characteristics influence health and development. More than 14000 mothers were enrolled during their pregnancy in 1991 and 1992, and the health and development of their children has been followed in great detail ever since (Golding et al., 2001). The study maintains a huge amount of genetic and phenotypic data, including genetic, epigenetic, biological, psychological, social and other environmental exposures from across participant life course (Boyd et al., 2013). In this work, 9233 ALSPAC individuals were genotyped using the Illumina HumanHap550 quad genome-wide SNP genotyping platform (Illumina, San Diego).

ALSPAC individual-level genotyped data will be used as a reference panel for fine mapping atopic dermatitis associated loci in Chapter 6.

1.5.4. Meta-analysis of cardiovascular disease traits

The University College-London School-Edinburgh-Bristol (UCLEB) consortium is comprised of a set of population-based prospective studies with dense CVD related phenotypes ((Chapter 9.13 UCLEB)). All participating studies have genotyped their individuals using a new high-resolution custom SNP array (“Metabochip”, see Section 2.1.1). The array have a much higher SNP densities in genomic regions linked to cardiovascular and metabolic diseases / traits.

The overall aims of the UCLEB consortium include, a) follow-up studies of genomic regions reported to be associated with CVD traits by previously GWAS studies and understand the unknown mechanisms for healthcare; b) to fine map functionally relevant SNPs inside these regions; c) to estimate the CVD risks for individuals and population separately based on individual SNPs and their interactions; d) to understand the mechanisms leading to altered CVD risk factor profiles; e) to process Mendelian Randomisation to identify the causal role in CVD of a range of cardiovascular biomarkers to inform public health policy and help develop new preventative therapies, i.e. new drug targets (Shah et al., 2013b).

1.5.5. Meta-analysis of gallbladder disease

The meta-analysis of gallbladder disease is a collaboration based project between Candidate gene association resource (CARE) consortium (Musunuru et al., 2010a) (Chapter 9.13 CARE) and UCLEB consortium. It consists of three independent cohorts: the Atherosclerosis Risk in Communities Studies (ARIC), BWHHS and Women's Health Initiative (WHI). ARIC is a prospective epidemiology study designed for atherosclerosis and cardiovascular disease (CVD) researches. In total, it has recruited 15,792 individuals between 1987 and 1989 (The ARIC investigators, 1989). The WHI is a National Institutes of Health (NIH) study and has recruited 161,808 women with ages between 50 and 79 years. The aim of WHI study is to prevent diseases (i.e. heart diseases) in women (The Women's Health Initiative Study Group, 1998). Participants from these three cohorts were genotyped using the "Cardiochip" (see Section 2.1.1)

1.5.6. Meta-analysis of Atopic Dermatitis

The EARly Genetics and Lifecourse Epidemiology (EAGLE) Consortium is a collaboration between various cohorts and aims to investigate the genetic basis of a range of phenotypes in antenatal and early life and childhood (Paternoster et al., 2012). The participating cohorts within the EAGLE consortium includes: the Australian Asthma Genetics Consortium (AAGC) replication cohort, the Avon Longitudinal Study of Parents and Children (ALSPAC), BAMSE, British 1958 birth cohort (B58C), the Children's Hospital of Philadelphia (CHOP), the Copenhagen

Prospective Study on Asthma in Childhood (COPSAC), Danish National Birth Cohort (DNBC), the European Community Respiratory Health Survey (ECRHS), Generation R, Genetics of Overweight Young Adults (GOYA) women's study, the Danish Glostrup Cohort (Health2006), KORA, LISA/GINI, Manchester Asthma and Allergy Study (MAAS), the Norwegian Mother and Child Cohort Study (MoBa), the Northern Finland Birth Cohort 1966 (NFBC66), Northern Finland Birth Cohort 1986 (NFBC86), Netherlands Twin Register (NTR), PIAMA, QIMR discovery cohort, Western Australian Pregnancy (Raine) cohort, SAPALDIA, the Department of Twin Research and Genetic Epidemiology at King's College London (TwinsUK). Meta-analysis summary results of the EAGLE consortium of Atopic Dermatitis were conducted by Paternoster and colleagues. I will use them together with the ALSPAC genotype data to run a regional association analysis in Chapter 6.

1.5.7. Meta-analysis of Body Mass Index

The Genetic Investigation of ANthropometric Traits (GIANT) consortium is an international collaboration based consortium (http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium) (Chapter 9.13 GIANT). It aims to identify genetic loci that modulate human body size and shape, including height and measures of obesity. The GIANT consortium consist of investigators from many different cohorts, institutions, countries, and studies, and the results represent their combined efforts (i.e. meta-analysis summary statistics). The primary analysis has been meta-analysis of GWAS data and other large-scale genetic data sets. Anthropometric traits that have been reported

by GIANT include body mass index (BMI), height, and traits related to waist circumference (such as waist-hip ratio adjusted for BMI, or WHRadjBMI). So far, the GIANT consortium has identified more than hundreds genomic regions with common genetic variants that are associated with anthropometric traits. The GIANT meta-analysis summary results of BMI (Speliotes et al., 2010) will be used for a further fine mapping analysis in Chapter 4.

1.6. Bioinformatics tools

1.6.1. Variants database: dbSNP

The National Centre for Biotechnology Information (NCBI) dbSNP database is a general catalogue of genetic variation (mainly single nucleotide variations and small INDELs) which is open access to the public (Sherry et al., 1999) (Sherry et al., 2001). The web link of dbSNP database is <http://www.ncbi.nlm.nih.gov/projects/SNP/>. The dbSNP database is the central repository for genetic variation and contains both polymorphisms and clinical mutations. At the time of writing, the most up-to-date database (dbSNP Build 140) contained ~249 million variant submissions (ss#s), of which ~161 million are unique positions in the genome (rs#s) and ~45 million have been validated (data obtained from http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi).

Day et al summarized that data quality is a big issue for dbSNP since various studies provide evidence for a high number of false positive ‘multisite’ variants which arise because of genotyping errors associated with paralogous (duplicated) genes (Day,

2010). Although the number of submissions to dbSNP has increased over the last few years, it is estimated that up to 8.3% of the unique coding variants could be false positives (Musumeci et al., 2010).

1.6.2. SNP Annotation tools: Variant Effect Predictor (VEP) and SNP Annotation and Proxy Search (SNAP)

Variant Effect Predictor (VEP)

Variant Effect Predictor (VEP) is a tool to predict the consequences of genomic variants and annotate variants onto known transcripts (McLaren et al., 2010). The VEP web interface is: http://www.ensembl.org/Homo_sapiens/Tools/VEP

VEP will be used to annotate variants significantly associated with Atopic Dermatitis in Chapter 6.

SNP Annotation and Proxy Search (SNAP)

SNP Annotation and Proxy Search (SNAP) is a bioinformatics query tool for SNPs to identify and annotate nearby SNPs in LD (called proxies) based on HapMap and 1000 Genome (Johnson et al., 2008). It can be used to find proxy SNPs as well as pairwise LD information. Moreover, it can be used to visualize the regional association analysis among variants within a genomic region.

1.6.3. Batch Coordinate Conversion (LiftOver)

Batch coordinate conversion (liftOver) converts genome coordinates and genome annotation files between assemblies (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>).

To conduct a LiftOver, we need an input file called BED file. An example of the BED file is shown in Table 1-5.

Chr	Start POS	End POS	Name
chr7	127471196	127472363	Region 1
chr7	127472363	127473530	Region 2
chr7	127473530	127474697	Region 3
chr7	127474697	127475864	Region 4
chr7	127475864	127477031	Region 5
chr7	127477031	127478198	Region 6
chr7	127478198	127479365	Region 7
chr7	127479365	127480532	Region 8
chr7	127480532	127481699	Region 9

Table 1-5. *Example of BED file.*

In addition, a chain file that links different assemblies (i.e. hg18 and hg19) is needed. Both files can be downloaded from the UCSC webpage. The code of LiftOver were shown in Chapter 9.1.1.

1.6.4. Linkage disequilibrium tools: Haploview, CubeX and MIDAS

Haploview

Haploview is a tool developed by the Broad Institute (Barrett et al., 2005). It is designed for analysing and visualizing haplotypes and LD. As illustrated in Figure 1-10, the interface of Haploview is simple and user friendly. In this work, I used Haploview as a tool to draw linkage disequilibrium (LD) plots. The input of Haploview can be either HapMap haplotypes data or our own data in a PLINK input format (e.g. PED and MAP file).

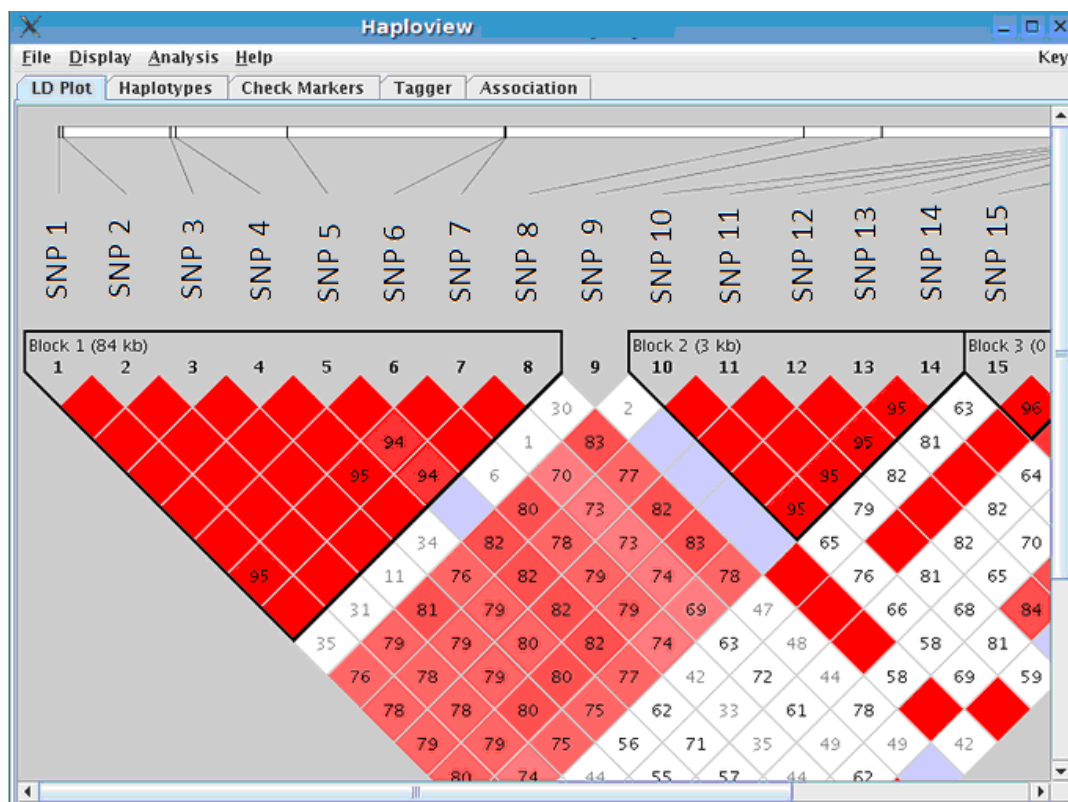
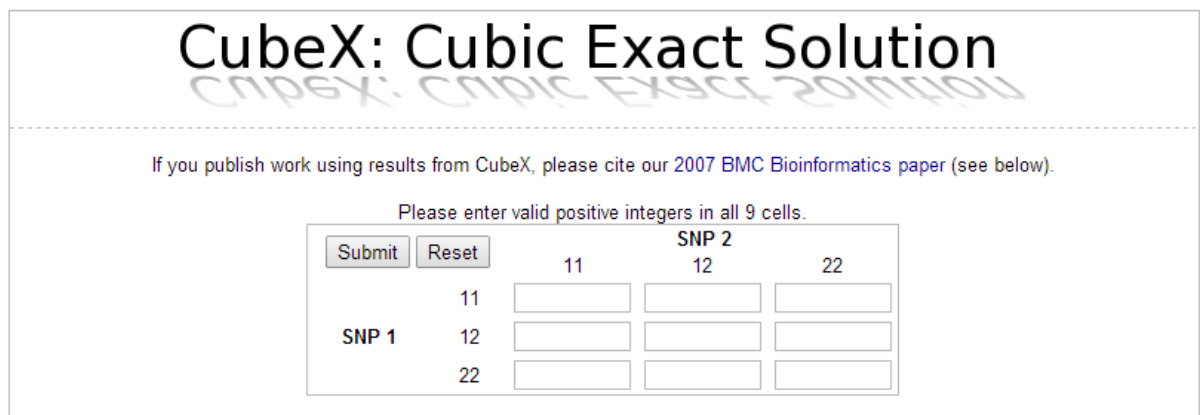


Figure 1-10. Haplotype interface and example of a LD plot.

CubeX

CubeX is a tool to estimate the pairwise haplotype frequencies using cubic exact solutions (Gaunt et al., 2007). As shown in Figure 1-11, the inputs to CubeX are the number of individuals of each genotype for the 3x3 genotype table. The output will be all possible haplotype frequencies for the two test SNPs. CubeX will be used in Chapter 3 to calculate the haplotype frequencies (X11) from reference genotype panel (i.e. HapMap and BWHHS individual-level genotype data).



The screenshot shows the CubeX web interface. At the top, the title "CubeX: Cubic Exact Solution" is displayed in a large, bold font. Below the title, a dashed line separates the header from the main content area. The main content area contains a message: "If you publish work using results from CubeX, please cite our 2007 BMC Bioinformatics paper (see below)." followed by a link. Below this, a instruction "Please enter valid positive integers in all 9 cells." is shown. The input form consists of a 3x3 grid of text boxes. The columns are labeled "SNP 2" with values 11, 12, and 22. The rows are labeled "SNP 1" with values 11, 12, and 22. To the left of the grid are two buttons: "Submit" and "Reset".

		SNP 2		
		11	12	22
SNP 1	11	<input type="text"/>	<input type="text"/>	<input type="text"/>
	12	<input type="text"/>	<input type="text"/>	<input type="text"/>
	22	<input type="text"/>	<input type="text"/>	<input type="text"/>

Figure 1-11. Web-based tool of CubeX.

MIDAS

Multiallelic Interallelic Disequilibrium Analysis Software (MIDAS) is a software for LD analysis with a comprehensive graphical users interface. It illustrates the patterns of LD between all types of multiallelic and biallelic markers (Gaunt et al., 2006) (Figure 1-12). In this work, MIDAS will be used in Chapter 3 and Chapter 4 to represent the LD patterns of pair-wise LD.

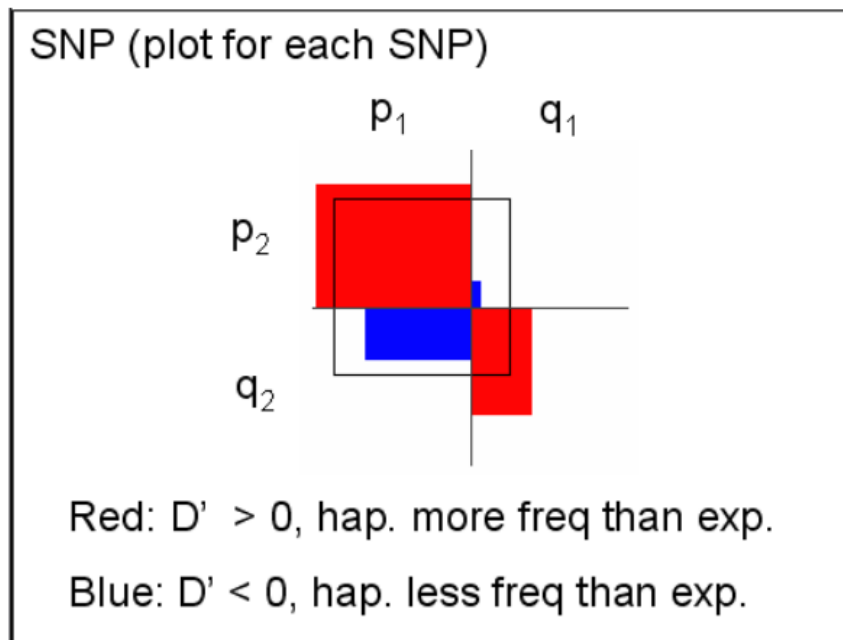


Figure 1-12. *Example of MIDAS results plot*

1.6.5. Functional prediction tools: FATHMM and RegulomeDB

Functional Analysis through Hidden Markov Models (FATHMM)

The Functional Analysis through Hidden Markov Models (FATHMM) is a high-throughput algorithm capable of predicting the functional, molecular and phenotypic consequences of protein missense variants using hidden Markov models (HMMs) representing the alignment of homologous sequences and conserved protein domains (Shihab et al., 2013). In this work, FATHMM will be used in Chapter 7 for functional prediction of coding SNPs associated with QTc interval.

RegulomeDB

RegulomeDB is an approach and database that guides interpretation of regulatory variants in the human genome. RegulomeDB uses high-throughput, experimental data sets from ENCODE (ENCODE Project Consortium et al., 2012) and other sources. It also includes computational predictions and manual annotations to identify putative regulatory potential and to identify putative functional variants (Boyle et al., 2012). The RegulomeDB scoring system was listed in Table 1-6. RegulomeDB will be used in Chapter 7 for functional prediction of non-coding SNPs associated with QTc interval.

Score	Supporting data
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak
2b	TF binding + any motif + DNase Footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
4	TF binding + DNase peak
5	TF binding or DNase peak
6	other

Table 1-6. *The RegulomeDB score system.*

1.7. Statistical genetics tools and programming languages

1.7.1. PLINK

PLINK is an open source C/C++ tool set for handling and analysing GWAS data in computational efficient manner (Purcell et al., 2007). The original functions such as data management (PED/MAP files, BED/BIM/FAM files and data filtering), population stratification (principal component analysis), summary statistics (estimating allele frequencies) and association analysis (for both binary traits and quantitative traits) will be used in the next few chapters. More importantly, I will use some of the updated functions included in version 1.07, such as LD calculation (estimating LD block, calculating LD measures D' and r^2), multimer haplotype analysis (haplotype phasing) and conditional analysis (for fine mapping individual-level data). I will use these functions for my novel methods development in Chapter 3 and Chapter 5.

1.7.2. GCTA conditional and joint SNP effect analysis

Genome-wide Complex Traits Analysis (GCTA) is a multiple function tools originally designed to estimate the proportion of phenotypic variance explained by genome wide SNPs for complex traits (Yang et al., 2011). It has now been extended for many other analyses to understand the genetic architecture of complex traits. In this work, I will focus on its conditional and joint genome-wide association analysis,

which can use summary-level statistics from a meta-analysis of genome-wide association studies (GWAS) and estimate linkage disequilibrium (LD) from a reference sample with individual-level genotype data. The command and options of the conditional and joint SNP effect analysis can be found in the following webpage: <http://www.complextaitgenomics.com/software/gcta/massoc.html>

1.7.3. R and Python

R is an implementation of the S programming language and is a free software environment for both high-level programming, statistical analysis and plotting graphics (<http://www.r-project.org/>). With the help of Rstudio (<https://www.rstudio.com/>), it has now become one of the most popular and user friendly language all over the world. In this work, R will be used in four respects: data managing, graphics plotting, processing statistical analyses and novel algorithm scripting.

On the other hand, Python is a widely used high-level programming language (<https://www.python.org/>). It is attractive since supports programming paradigms, including object-oriented, imperative and functional programming. Python can also be used with a wide range of third-party tools such as “matplotlib” and “RPy2”.

In this work, Python will be used in three respects: transforming data format, programming web-based tools and scripting a novel E-M algorithm.

1.8. Major aims and objectives

As I described in Section 1.4, applying existing variable selection methods, such as conditional analysis and model selection approaches, requires individual level data. Single GWAS lack power to obtain accurate association results (Zeggini and Ioannidis, 2009). Meta-analysis of multiple GWAS studies is a good way to increase the power of the association analysis. But pooling of individual genotype and phenotype data are normally not available from meta-analyses. As a consequence, the bulk of this thesis concerns the development of two novel approaches/algorithms so that meta-analysis results can be used directly for fine mapping analysis. Both methods will combine statistical concepts of regression analysis with population genetics concepts of allele frequencies, LD, haplotypes and haplotype frequencies. Furthermore, these two methods will be applied to real meta-analysis summary statistics to identify secondary signals associated with complex traits of human beings.

1.8.1. Objectives

- Use concepts of pair-wise LD and simple regression analysis to develop an approximate conditional analysis methods using meta-analysis summary data
- Investigate the performance of the approximate conditional analysis and compare it with conditional analysis and model selection methods

- Apply the approximate conditional analysis to cardiovascular disease (CVD) related meta-analysis to identify independent effect SNPs in the candidate regions
- Extend multiple regression analysis to group level and use E-M algorithm to develop an approximate joint SNP effect approach using meta-analysis summary statistics
- Perform an independent benchmark to compare the performance of the approximate joint SNP effect approach against other existing fine mapping methods.

Identify independent SNPs of 12 complex quantitative traits and 2 disease outcomes using the approximate joint SNP effect approach to identify more multiple SNPs associated with these traits / diseases.

CHAPTER 2. SEQUENTIAL SENTINEL SNP REGIONAL ASSOCIATION PLOTS USING META-ANALYSIS SUMMARY STATISTICS

2.1. Introduction

Genome wide association studies (GWAS) have been successful in identifying over twelve thousand SNPs involved in the development of human complex traits and disease outcomes (Hindorff et al., 2009) (Manolio, 2010). A considerably larger number of SNPs show significant associations (with respect to the genome-wide significance threshold) with these traits, but this can be caused either by the own effect of these SNPs or they are in LD with the causal variant (i.e. SNPs shown in Table 8-1). Moreover, with the increasing in sample size and the number of meta-analyses of multiple GWASs carried out, an increasing number of genetic variants associated with different traits will be identified consequently. With these in mind, fine mapping, which is an approach taken to identify secondary association signals at a certain genomic region, will become statistically feasible and a potentially powerful tool in the future. However, there is an untested assumption of GWAS which contradicts this idea and therefore slowed down the development of fine mapping approaches.

2.1.1. One Limitation of single SNP model GWAS, conditional analysis and model selection approaches

For GWAS, a single SNP model is usually used to test the association between phenotypes and genotypes. The most significant (noted as top-hit) SNP will represent the association of a genomic region such as a gene and/or a linkage disequilibrium (LD) block. In the following situations, this untested assumption may not be met: Firstly, this design does not take into account the possibility of multiple causal variants at a region. Secondly, the real causal variant maybe untyped or with low imputation quality; and the best genotyped (or imputed) SNP with LD to the real causal variant may not capture the total amount of variation at the region. Therefore, a GWAS based on a single SNP model will result in an underestimate of the total variation that could be explained at a locus (Yang et al., 2012).

To detect multiple potentially causal sites for GWAS, several statistical stepwise procedures exist for identify SNP independence using individual level data, which include statistical packages for conditional analysis (Purcell et al., 2007) and model selection criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallows' Cp, Cross-validation, adjusted r^2 and stepwise regression (Smith et al., 2010a). Among these methods, conditional analysis is the most commonly used tool and considered the gold standard for fine mapping when using individual-level data.

2.1.2. Meta-analysis and limitation of current meta-analysis system

As previously described in Chapter 1, meta-analysis of multiple GWASs has been shown to increase the power to detect genetic variants associated with complex disease risk traits such as blood lipids (Teslovich et al., 2010) (Asselbergs et al., 2012) and Electrocardiographic (ECG) traits (Arking et al., 2006) (Pfeufer et al., 2009) (Newton-Cheh et al., 2009) (Marroni et al., 2009) (Gaunt et al., 2012). The common meta-analysis design requires different research groups to manage individual level analyses of the same phenotype and then collaborate on detailed meta-analyses of this phenotype (Seminara et al., 2007). In fact, meta-analysis provides an improved platform for regional fine mapping with respect to GWAS.

Meta-analysis summary statistics

For a certain SNP, the most notable information one gets from a meta-analysis is the effect size of the SNP (noted as beta coefficient, beta or b), standard error of the effect size (SE) and the p value against the null hypothesis in regards to the SNP's association with the trait under analysis (p). Beta represents the overall effect of the SNP, where p value shows the level of significance of the SNP. On the other hand, the risk allele (A_2), alternative allele (A_1) and allele frequencies of the risk allele gives one the idea of the proportion of a particular allele within a population of N individuals.

Limitation of conventional meta-analysis system

Meta-analysis has proven to be a powerful method to identify disease susceptibility and quantitative trait loci which single GWAS did not identify or provide enough evidence due to lack of power (Ioannidis et al., 2006). As abovementioned, meta-analysis is based on collaboration, so data sharing is an issue amongst different institutes. Normally only standard summary statistics (as shown in Table 1-4) might be available for researchers needed data in the trait(s) of interest. Unfortunately, statistical approaches like conditional analysis and model selection methods are only applicable when individual-level data is available. Before I developed the novel method SSS-RAP (Zheng et al., 2013), if analysts from different institutes wanted to fine mapping a candidate region, they needed to processed conditional analysis in each single study and combined summary results together through a second round of meta-analysis. As we can imagine, it normally takes months for a single round of conditional analysis under the current meta-analysis system (Zheng et al., 2013).

To overcome the limitation of GWAS and increase the efficiency and accuracy of fine mapping under the current meta-analysis system, a biological point of view is necessary. Haplotypes, which contain all LD information for a certain genomic region, will be a good choice. Using rule of thumb “Simpler the better”, I started analyzing the simplest haplotypes, ones which contained two SNPs. Digging the correlations between pair-wise SNPs, especially correlation between top significant SNP and a query dependent SNP, became the first priority of this project.

2.1.3. Haplotype, haplotype frequency and r^2 measure of linkage disequilibrium from open access resources

For a meta-analysis combining data from a major ancestral group, for example, East Asian, the individual-level data will, most of the time, not be available. We are not able to estimate haplotype, haplotype frequencies and LD correlation (r^2) without individual level data. But through utilizing available resources such as the International HapMap Consortium (International HapMap 3 Consortium et al., 2010) and the 1000 Genome project (1000 Genomes Project Consortium et al., 2010) access to reliable individual level data can be possible. These open resources provide users with hundreds of well-defined references with individual-level phased haplotypes for each ancestral group. Previous studies have proven that a reference panel from the same population as the meta-analysis individuals will share similar LD correlation between a pair of SNPs (Ke et al., 2004). Under this assumption, two loci haplotype frequencies can be estimated freely using these data.

2.1.4. Limitation of considering r^2 as the standard of SNP independence

As discussed in chapter 3.1.2, fine mapping under the current meta-analysis system is time consuming. Since r^2 is a good representative of LD correlation and easy to calculate, some of the early meta-analysis papers used the LD correlation (r^2) as the

only standard to test SNP independence. But r^2 is a measure of pair wise LD, which cannot take into account the correlation among multiple SNPs. Besides, I hypothesized that, even in the two SNPs model, only considering the r^2 would not be adequate to detect the independent signals between the two loci. I used the following examples to prove the hypothesis: (i) Two SNPs in Low LD and (ii) Two SNPs in linkage equilibrium (which means there is no LD between the two SNPs).

Two SNPs in Low LD

In this section, I would prove that even if two SNPs are in low LD to each other, one can still be dependent to the other. Here I listed some real case example from two published papers (Gaunt et al., 2012) (Shah et al., 2013b), which, in my point of view, will be the best observational proof for my assumption. As illustrated in Table 2-1, I listed the conditional analysis p values of some pairs of SNPs reported in the above two papers. Amongst all these pairs, the conditional analysis suggested that these SNPs are dependent to each other although their r^2 is lower than 0.15

traits	gene	top hit	query	r^2	conditional
			dependent SNP		analysis p-value
QTc	<i>NOS1AP</i>	rs7534004	rs4657154	0.144	0.116
PR	<i>SCN5A</i>	rs12053903	rs7374540	0.172	0.198
QRS	<i>SCN5A</i>	rs7374540	rs11710077	0.113	0.166
HDL	<i>CETP</i>	rs3764261	rs12708967	0.111	0.066
TG	<i>LPL</i>	rs3735964	rs7016529	0.002	0.224
TG	<i>LPL</i>	rs3735964	rs1121923	0.002	0.608

Table 2-1. *Examples of query dependent SNPs with low LD ($r^2 < 0.223$) to the top hits but dependent on the top hits*

Two SNPs in linkage equilibrium (No LD)

I simulated one extreme case of two SNPs in linkage equilibrium. In this case, I assumed that SNP1 (allele frequency of minor allele=0.3) and SNP2 (allele frequency of minor allele =0.4) were two significant SNPs associated with a trait and there is no LD between them (Figure 2-1 a). Then I estimated the haplotype frequencies between SNP1 and SNP2 (Figure 2-1 b). Then I introduced a hidden SNP, SNP3, with allele frequency of minor allele equal to A or \bar{A} . Since the haplotype of SNP3 is known, I listed the haplotype frequencies between SNP1~SNP3 and SNP1~SNP2 in Figure 2-1c and d. Using these information, I calculated A , \bar{A} , D and D -bar. Interestingly, both A and \bar{A} was 0.28. This means when allele frequency of minor allele of SNP3 is 0.28, this SNP will be in LD to both SNP1 and SNP2 although there is no LD between SNP1 and SNP2.

This special case proves that even if two SNPs are in linkage equilibrium, both of them can still be in LD with a third SNP. If the third SNP is the untested causal signal, only considering the LD between the two tested, significant SNPs will be problematic.

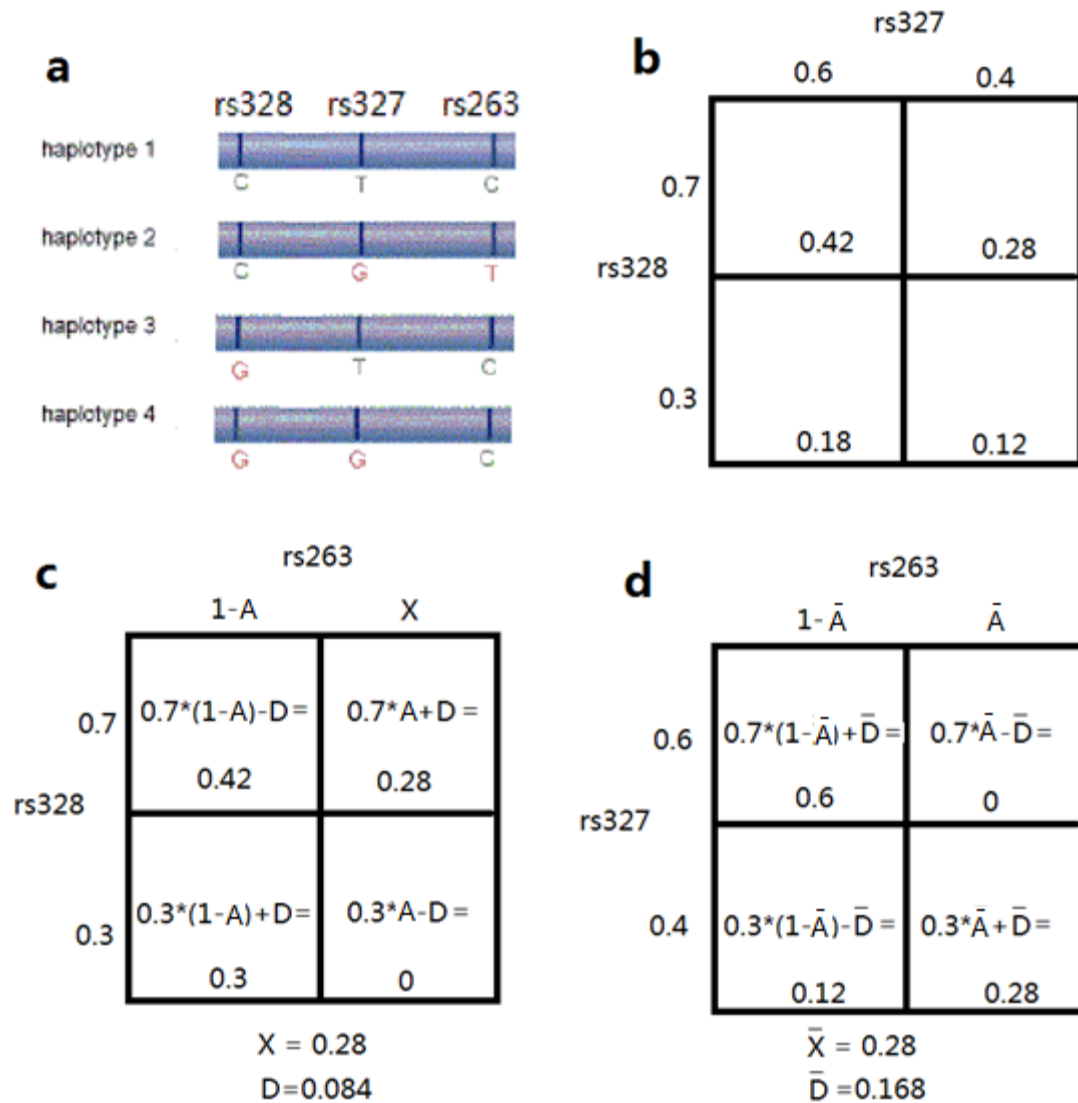


Figure 2-1. A special case using the three SNPs model. **a**, the three SNPs model; **b**, rs327 and rs328 are in linkage equilibrium; **c** and **d**, estimating allele frequencies for rs263 with certain haplotype frequencies. Values at the bottom of each grid are the haplotype frequencies. D and D bar were the LD measures between rs263 and rs327/rs328

2.1.5. Aims and objectives

At the time I wrote this chapter, the step-wise regression analysis (Efroymson, 1960)-the most widely used method for identify SNP independence-required

individual-level genotype data, which is normally not achievable in the current meta-analysis system (Seminara et al, 2007). On the other hand, I also proved that only consider LD relationship (r^2) is not a ideal way of identifying SNP independence. Before the end of 2012, this is a blank area in genetic fine mapping field, which worth digging more. So in this chapter, I shall describe the development of a novel fine mapping method, which shall only use meta-analysis summary statistics. It should be highly automated and computational fast. The method is called sequential sentinel SNP regional association plots (SSSRAP, pronounced (es,es,es-rap)), which is a step-wise elimination method to detect the independent effect of query dependent SNPs conditional on the top significant signals. The dependence between top signal and the test SNPs shall be shown be depicted in a plot. The method shall be applicable to group-level data as well as individual-level data. It shall only require meta-analysis summary information and the haplotype frequencies information estimated from a reference panel. In addition, different functions shall be developed so that the method can be applicable to both quantitative (i.e. continuous) traits and binary (i.e. disease) outcomes.

Objectives

- Develop a novel fine mapping method using meta-analysis summary statistics of complex quantitative / disease traits
- Investigate how the standard error of beta depends on minor allele frequency

- Compare the performance of the program against the performance of ten existing fine mapping methods using samples from 1) BWHHS individual-level data, 2) 1000 Genome individual-level data and 3) several simulated datasets
- Develop a user friendly web-based tool for the novel method

2.2. Methods and Materials

2.2.1. Formulation of problems: linking meta-analysis summary statistics to pair-wise linkage disequilibrium measures

For a pair of SNPs significant associated with a complex trait, for example lipids, it is not possible to detect whether the two SNPs are independent of each other using the meta-analysis summary statistics (single SNP effects) directly. As illustrated in Figure 2-2, the haplotype and haplotype frequencies of two SNPs can be used to link the observed beta (marginal SNP effect) of the top hit (top significant SNP) to the dependent effect (noted as the transformed beta) of the query dependent SNP by using a 2x2 table transformation. Then for the query dependent SNP, I compared the observed beta and its transformed beta. If two betas are statistically different, then I concluded that the query dependent SNP is independent conditional on the top hit and *vice versa*. The assumptions and notations of the SSSRAP program were listed in Table 2-2. The flow chart of the whole SSSRAP program was illustrated in Figure 2-4.

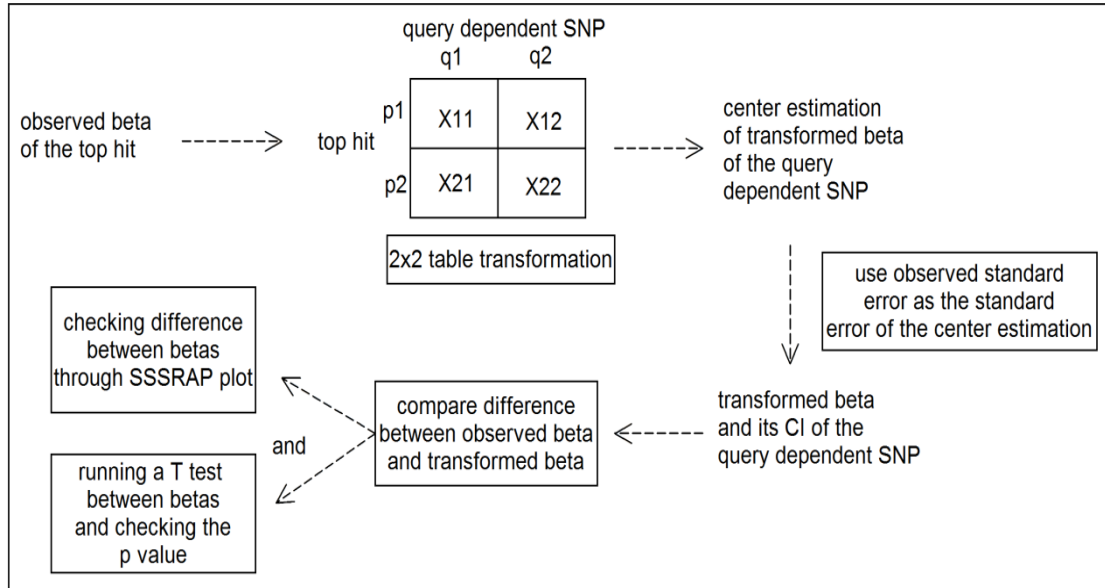


Figure 2-2. Two SNPs 2x2 haplotype table and regression coefficient transformation

Assumption:

- Linkage disequilibrium between pair wise SNPs (or it will be meaningless to test the SNP independence)
- For disease outcomes, the log odds increase equally with each addition of the minor allele (for each SNP)
- For quantitative traits, the difference of effect between homozygous wild type genotype and heterozygous genotype is equal to the difference of effect between heterozygous genotype and homozygous mutant genotype
- Participants are from the same population in the meta-analysis I applied
- The top hit selected for the first phase tags a single causal signal.
- Although the observed beta and transformed beta are not totally independent of each other, I still use the two samples unpaired T test.
- Assume normal distribution of the phenotype, so I need to do the normalization of the phenotypes

Definitions:

p_1 and p_2 represent the major and minor allele frequency of the top significant SNP

q_1 and q_2 are the major and minor allele frequency of the possible dependent SNP

X_{11} , X_{12} , X_{21} , X_{22} represent the four haplotype frequencies for the two respective SNP

loci

$Odds_{maj}$ and $Odds_{min}$ represent the disease odds of each major and minor allele respectively

'e' and 'b' represent the effect per minor allele and per major allele for the top hit respectively.

β_{obs_top} is the observed beta of the top hit

β_{trans} is the beta estimation of the possible dependent SNP

$Odds_{maj_top}$ and $Odds_{min_top}$ are the odds of major and minor allele for the top hit respectively

$Odds_{maj_trans}$ and $Odds_{min_trans}$ are the odds of major and minor allele for the possible dependent SNP respectively

'm' is the number of SNPs in the test group

SE_{obs} is the observed standard error for each possible dependent SNP

SE_{trans} is the transformed standard error for each possible dependent SNP

SE_{sim} is the simulated standard error for each possible dependent SNP

Notes:

X_{11} should be smaller than both p_1 and q_1

Table 2-2. *Assumptions and notations*

2.2.2. Relating regression coefficient to minor allele frequencies and haplotype frequencies

As explained in Chapter 1, in the association analysis field, regression coefficient (beta) is the core value of the genetic association analysis, which is the gradient of the regression line of a single SNP model analysis.

For disease outcomes, the regression coefficient for a specific SNP is the natural logarithm of the odds ratio:

$$\beta = \text{LN (Odds Ratio)} = \text{LN} \left(\frac{\text{Odds of minor allele}}{\text{Odds of major allele}} \right) \quad (1)$$

As illustrated in Figure 2-2, for a pair of SNPs, the observed beta of the top hit, noted as $\beta_{\text{obs_top}}$, can be transformed into the estimated beta of the query dependent SNP, noted as β_{trans} , through the 2x2 table transformation.

Using equation 1, the $\beta_{\text{obs_top}}$ and β_{trans} can be extended to:

$$\beta_{\text{obs_top}} = \text{LN}(\text{Odds}_{\text{min_top}} / \text{Odds}_{\text{maj_top}}) \quad (2)$$

$$\beta_{\text{trans}} = \text{LN}(\text{Odds}_{\text{min_trans}} / \text{Odds}_{\text{maj_trans}}) \quad (3)$$

Where $\text{Odds}_{\text{maj_top}}$ and $\text{Odds}_{\text{min_top}}$ are the odds of major and minor allele for the top hit respectively; $\text{Odds}_{\text{maj_trans}}$ and $\text{Odds}_{\text{min_trans}}$ are the odds of major and minor allele for the query dependent SNP respectively.

The above odds of alleles can be represented by the four haplotype frequencies X_{11} , X_{12} , X_{21} and X_{22} :

$$\beta_{\text{trans}} = \text{LN} \left(\frac{X_{22} * \text{Odds}_{\text{min_top}} + X_{12} * \text{Odds}_{\text{maj_top}}}{X_{22} + X_{12}} / \frac{X_{21} * \text{Odds}_{\text{min_top}} + X_{11} * \text{Odds}_{\text{maj_top}}}{X_{21} + X_{11}} \right) \quad (4)$$

The haplotype frequencies X_{22} , X_{12} and X_{21} can be represented as a combination of the allele frequencies p_1 , p_2 , q_1 and q_2 and haplotype frequency X_{11} :

$$\beta_{\text{trans}} = \text{LN} \left(\frac{(p_2 q_2 + X_{11} - p_1 q_1) * e^{\beta_{\text{obs_top}} + p_1 q_2 - X_{11} + p_1 q_1}}{q_2} / \frac{(p_2 q_1 - X_{11} + p_1 q_1) * e^{\beta_{\text{obs_top}} + X_{11}}}{q_1} \right) \quad (5)$$

For quantitative traits, beta of a locus is the slope of the regression line. For a pair of SNPs, the top hit beta is:

$$\beta_{\text{obs_top}} = e - b$$

Where e represents the effect per allele of homozygous mutant genotype and b represents the effect per allele of homozygous wild type genotype (Figure 2-3). Assume the effect of homozygous mutant genotype (AA) is $2e$, meanwhile the effect of homozygous wild type genotype (TT) is $2b$. The slope of the regression line (beta) will be $e-b$.

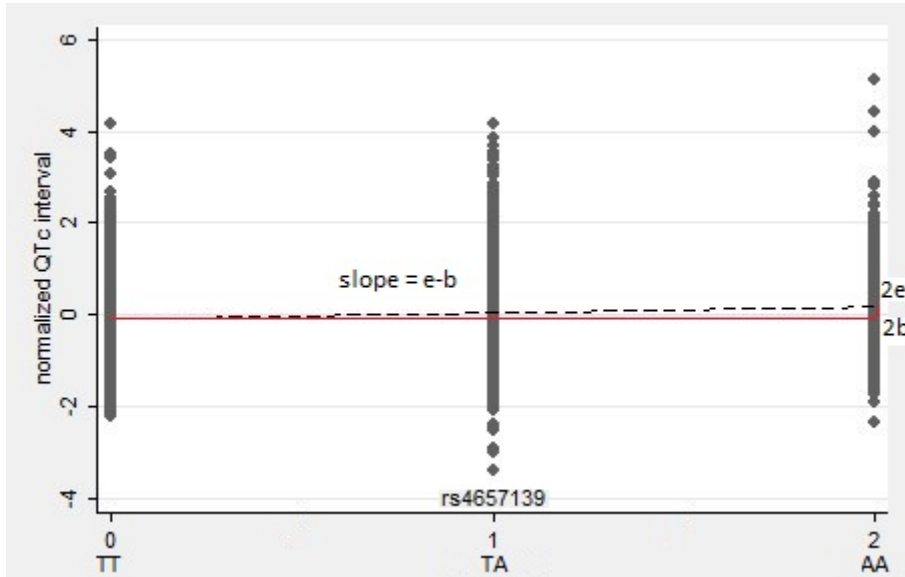


Figure 2-3. Example of a linear trend line for a single locus. The grey dots are phenotype values (QTc interval) for 2686 participations, the black dashed line is the linear trend line, the x axis is the genotypes of SNP rs4657139, y-axis is the normalized QTc interval (normalized by Z score).

As illustrated in Figure 2-2, the beta transformation formula becomes:

$$\beta_{\text{trans}} = \frac{e * N * X_{22} + b * N * X_{12}}{N * X_{22} + N * X_{12}} - \frac{e * N * X_{21} + b * N * X_{11}}{N * X_{21} + N * X_{11}} \quad (6)$$

Where N is the number of participants, X_{11} , X_{12} , X_{21} and X_{22} are the four haplotype frequencies for this pair of SNPs.

N cancels, and representing X_{22} , X_{12} and X_{21} using p_1 , q_1 and X_{11} , gives:

$$\beta_{\text{trans}} = \beta_{\text{obs_top}} * \left(\frac{x_{11}-p_1}{1-q_1} + \frac{x_{11}}{q_1} \right) \quad (7)$$

2.2.3. Estimating standard errors of transformed betas

To estimate the standard error of the transformed beta for each query dependent SNP, I separated the total uncertainty (where is the reason of the beta getting a specific standard errors) to two parts: phenotype uncertainty and haplotype uncertainty. The phenotype uncertainty is the errors caused by the randomness of the phenotype. The haplotype uncertainty, as illustrated in Chapter 1, is the proportion of the middle cell haplotype phase is unknown in a 3×3 table for a biallelic marker (Table 1-2). I assessed both phenotype uncertainty and haplotype uncertainty separately using a simulated dataset. The dataset was a group of SNPs that significantly associated with three ECG traits and three lipid traits in BWHHS (Table 8-1A and 9-1B). These data was chosen since previous reported multiple association signals in these regions, and these regions are genotyped using a high-throughput chip. The phenotypes of these traits were standardized using a Z score, which made the unit of the betas the same among these traits, so that I can combine them together for this simulation. For the phenotype uncertainty, I transformed the 95% CI of observed beta for each possible dependent SNP through the 2×2 contingency table directly and get the “95% confidence limit” of transformed beta (noted as SE_{trans}). To explain the haplotype uncertainty of the transformed beta, standard errors (noted as SE_{sim}) were derived from a distribution generated by 10,000 simulations.

2.2.4. Two tailed t test and stepwise elimination

Extending the pair-wise SNPs situation to a model containing 'm' loci, I estimate the transformed beta (different formula for disease outcomes or quantitative traits) for each pair of top significant SNP and query dependent SNPs. So I will derive 'm-1' transformed betas (from β_{trans_1} , to β_{trans_m-1}).

For each query dependent SNP "i" ($i = 1 \dots m-1$), the difference between its observed beta (β_{obs_i}) and its transformed beta (β_{trans_i}) will be quantified using an unpaired two tailed t test. The reason of choosing the two tailed t test is that the effect of SNPs can be either risk increasing or protective.

As the estimation results suggested in Results section and Figure 2-8, the observed standard error for the query dependent SNP (SE_{obs_i}) will be used instead of the standard error of its beta estimation (SE_{trans_i}). The t statistics of this two tailed t test becomes:

$$T_i = \frac{\beta_{obs_i} - \beta_{trans_i}}{\sqrt{SE_{obs_i}^2 + SE_{obs_i}^2}}$$

The t-statistics will then be transformed to t test p values using R code:

$$2 \times pt(-abs(t\text{-statistics}), df = N-1),$$

Then a step-wise elimination procedure is used to remove SNPs highly dependent on the top hit. The p-value threshold will be set (for example: 0.01) and SNPs with t test p value higher than the threshold will be removed from the model. For the remaining SNPs, the SNP with the lowest t test p-value will be used as the second

independent SNP and it will be used as the top hit for the next phase step-wise elimination (Figure 2-4).

Although the remaining SNPs are significant conditional on the top hit, the top hit may still have minor effect on them. To remove the effects of the top hit from the next phase elimination, I adjust the “observed” beta as follows:

$$\beta_{\text{obs_adjusted}}^{(g+1)} = \beta_{\text{obs}}^{(g)} - \beta_{\text{trans}}^{(g)}$$

Where $\beta_{\text{obs_adjusted}}$, β_{obs} and β_{trans} are adjusted observed beta, unadjusted observed beta and transformed beta of the SNP. “g” is the number of current phase of elimination, whereas g+1 is the next phase of elimination

The step-wise elimination will be continued until all SNPs with independent effects are detected (Figure 2-4).

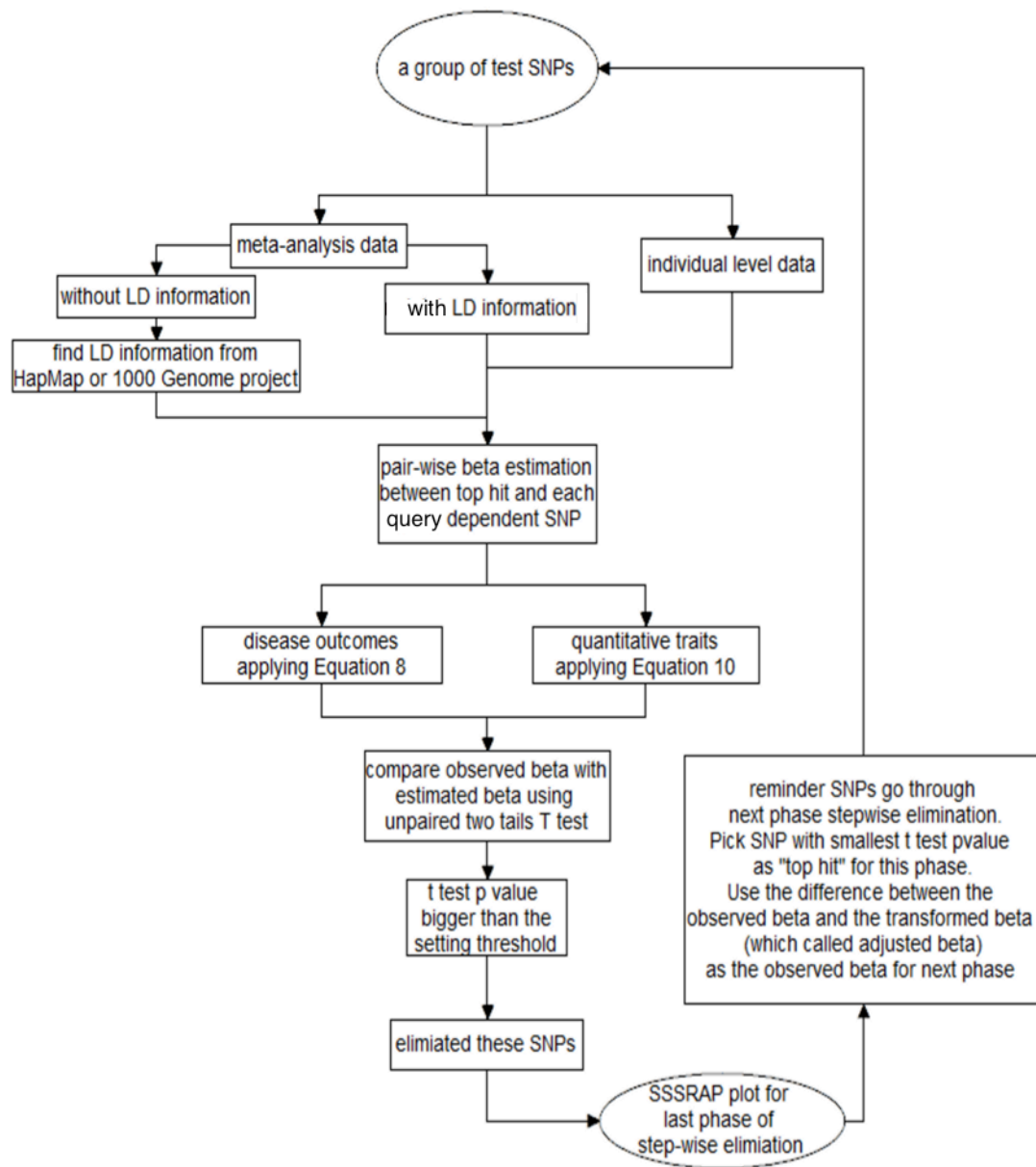


Figure 2-4. Flowchart of the SSSRAP approach

2.2.5. SSS-RAP web-based tool

A web-based implementation of the SSS-RAP program was developed using Python CGI (for details please see Appendix, Chapter 9.3). I illustrated the homepage in Figure 2-5. The URL is: <http://apps.biocompute.org.uk/sssrp/sssrp.cgi>



Figure 2-5. Homepage of SSS-RAP programme.

2.2.6. Sample datasets

The BWHHS sample data

The SSS-RAP program was applied to the analysis of BWHHS individual level data (Gaunt et al., 2012) (Shah et al., 2013a) for three ECG traits (PR interval, QRS duration and QTc interval) and two lipid traits (triglycerides and HDL cholesterol). SNPs selected for the performance comparison were listed in Table 8-1A and 9-1B.

To prepare the input data required, I controlled the quality of data for individuals with

missing genotypes and phenotypes using PLINK code `--geno` and `--prune`. I standardized phenotype using Z score, which is explained in Chapter 1. Using PLINK, I calculated MAF, conducted an association test using the additive SNP model and estimated haplotype frequencies for each pair of SNPs.

Using model selection methods, I aim to identify one subset of SNPs providing the best fit to the data among all the possible models. The model selection methods were selected using a number of criteria in R (<http://www.r-project.org/>, 2.14.1), which included Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallows C_p , Cross Validation, leave-one-out-cross-validation (LOOCV), adjusted r^2 , stepwiseAIC, stepwiseBIC, stepwise C_p , stepwise regression with p value threshold of 0.05. The concepts of these methods were explained in Chapter 1. I listed the data and R code I used in Chapter 9.1.2. Besides, the conditional analyses were run in PLINK with a p-value cut-off value of 0.001.

Admixed population simulation using 1000 Genome data

To analyse whether minor departures of LD (e.g. two sub-populations in one major ethnic group) affect the SSS-RAP assumption of one ancestral group, I combined the genotype data of HapMap unrelated Utah Residents (CEPH) with Northern and Western European ancestry (CEU) unrelated individuals and the Toscana in Italia (TSI) individuals. SNPs associated with PR interval and QRS duration in the ECG meta-analysis were selected (Gaunt et al., 2012). I applied SSS-RAP to the summary statistics (betas and SEs) of the meta-analysis and allele frequencies and haplotype frequencies estimated using individual-level genotypes of CEU individuals as well as the admixed population of CEU and TSI. Then I compared the results of these two designs.

Simulation to test relationship between r^2 and beta

I conducted several simulated datasets to test the relationship between r^2 and beta. For the top hit, I assumed that it has a MAF of 0.1. To test whether the method performed differently with different marginal SNP effects, I set three different marginal effects for the top hit: 0.5, 1 and 4. For the query dependent SNP, I varied allele frequencies and r^2 , increasing 0.01 per time in the range of 0.5 to 1 for allele frequencies and in the range of 0 to 1 for r^2 .

Admixed population simulation

For the BWHHS genotype data of LPL gene, I imputed the region using IMPUTE2. The HapMap genotypes were used as a reference panel for the imputation. I selected three SNPs with varying associations with triglyceride (TG): rs328, which is a well replicated top significant signal associated with TG; rs327, which is a significant SNP associated with TG and rs263, which is a bystander SNP not associated with TG. For these three SNPs, I used the four commonest haplotypes among them (Figure 2-6). We assume the frequencies of these four haplotypes change 0.1 each time. Haplotype 1 is the commonest haplotypes with frequency change between 0.5 to 0.8. The frequencies of the other three haplotypes are changing between 0.1 to 0.4. Totally there are 55 possible haplotype frequencies combination. Amongst these haplotype frequencies combinations, 20 were not in perfect LD (Table 2-3). These 20 combinations were used in this simulation. I simulated the genotypes for 10000 individuals for each set of these 20 haplotype frequencies combinations. The phenotypes for the 10000 individuals were estimated using the true effect plus a random baseline effect. These 20 simulated populations were noted as “pop_1”. The whole process

was repeated again and the second 20 simulated populations were noted as “pop_2”. Then for both pop_1 and pop_2, I picked the first population and combined them together as an admixture population. Using the same way, I created 210 admixed populations.

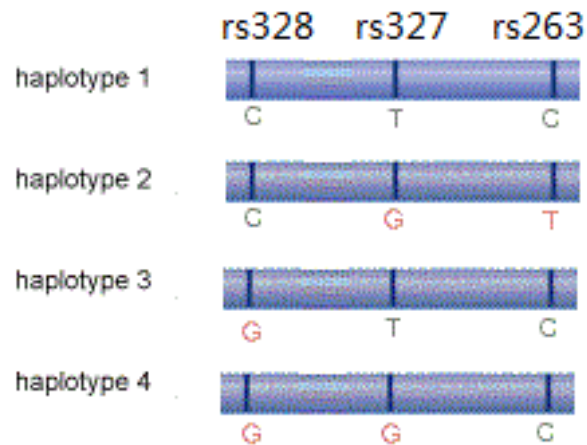


Figure 2-6. *The three SNPs model for the admixed population simulations.* I assumed rs328 was the top hit, rs327 was an independent effect SNP and rs263 was a by stand SNP. Alleles in red means they were the effect / minor alleles.

ID	haplotype 1	haplotype 2	haplotype 3	haplotype 4
5401	0.5	0.4	0	0.1
5311	0.5	0.3	0.1	0.1
5302	0.5	0.3	0	0.2
5221	0.5	0.2	0.2	0.1
5212	0.5	0.2	0.1	0.2
5203	0.5	0.2	0	0.3
5131	0.5	0.1	0.3	0.1
5122	0.5	0.1	0.2	0.2
5113	0.5	0.1	0.1	0.3
5104	0.5	0.1	0	0.4
6301	0.6	0.3	0	0.1
6211	0.6	0.2	0.1	0.1
6202	0.6	0.2	0	0.2
6121	0.6	0.1	0.2	0.1
6112	0.6	0.1	0.1	0.2
6103	0.6	0.1	0	0.3

7201	0.7	0.2	0	0.1
7111	0.7	0.1	0.1	0.1
7102	0.7	0.1	0	0.2
8101	0.8	0.1	0	0.1

Table 2-3. *Possible haplotype frequencies combination for the 3 SNPs model*

2.3. Results

2.3.1. Estimating standard errors of transformed betas

Previously in Chapter 3.2.3, I have explained my way of finding out the appropriate standard errors for the beta produced by SSS-RAP (noted as transformed beta). Here is the result of my hypothesis that minor allele frequency (MAF) has a major influence on the standard error of beta.

Relationship between minor allele frequencies and standard error of beta

As shown in Figure 2-7, with smaller numbers of individuals in the variant allele group (i.e. lower MAF), the errors (both observed SEs and simulated SEs) on estimates are wider, which can also be supported in a statistical way: the residual variance of the regression is a function of the inverse of the variance in X, which is a function of square of MAF. .

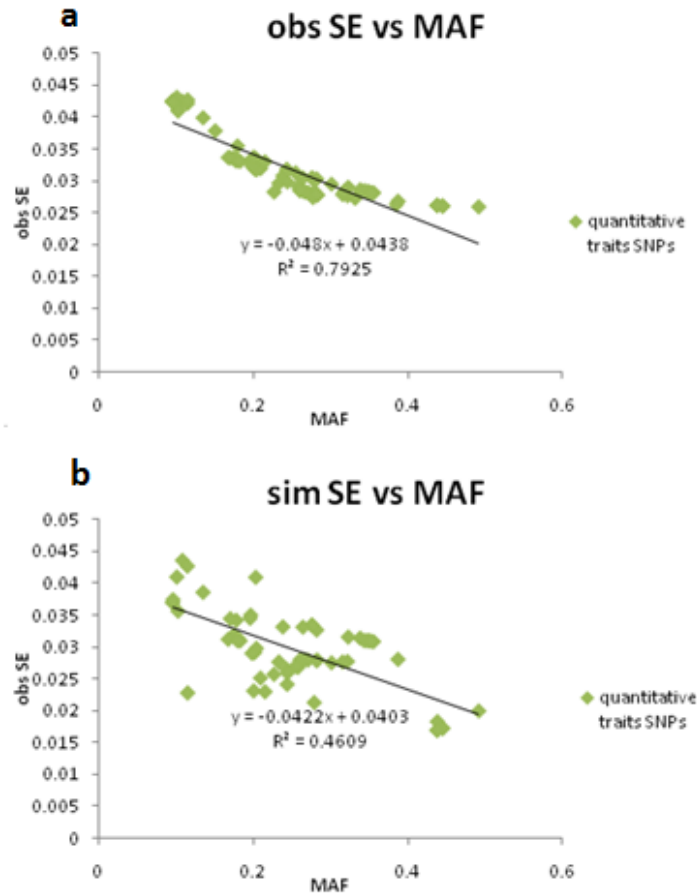


Figure 2-7. *The influence of minor allele frequency on the standard error of beta.*

a, relationship between the observed standard errors and the minor allele frequencies (MAF). **b**, relationship between the simulated standard errors and MAFs

Standard error of SSS-RAP betas

The total uncertainty (which is represented by the SE) of SSS-RAP beta was combination of the phenotypic uncertainty (transformed SE) and haplotypic uncertainty (simulated SE). I tested the relationship between the observed SE and both transformed SE and simulated SE in different r^2 ranges. When the r^2 is between 0 and 0.5, the observed SE of the possible dependent SNP was similar to the simulated SE and total SE (Figure 2-8a and 3-7b). In contrast, when r^2 is between 0.5 and 1, the observed SE was more similar to the

transformed SE and total SE (Figure 2-8e and 3-7f). So for the query dependent SNP, the observed SE of the single SNP regression is an appropriate representative of the SE of the SSS-RAP beta.

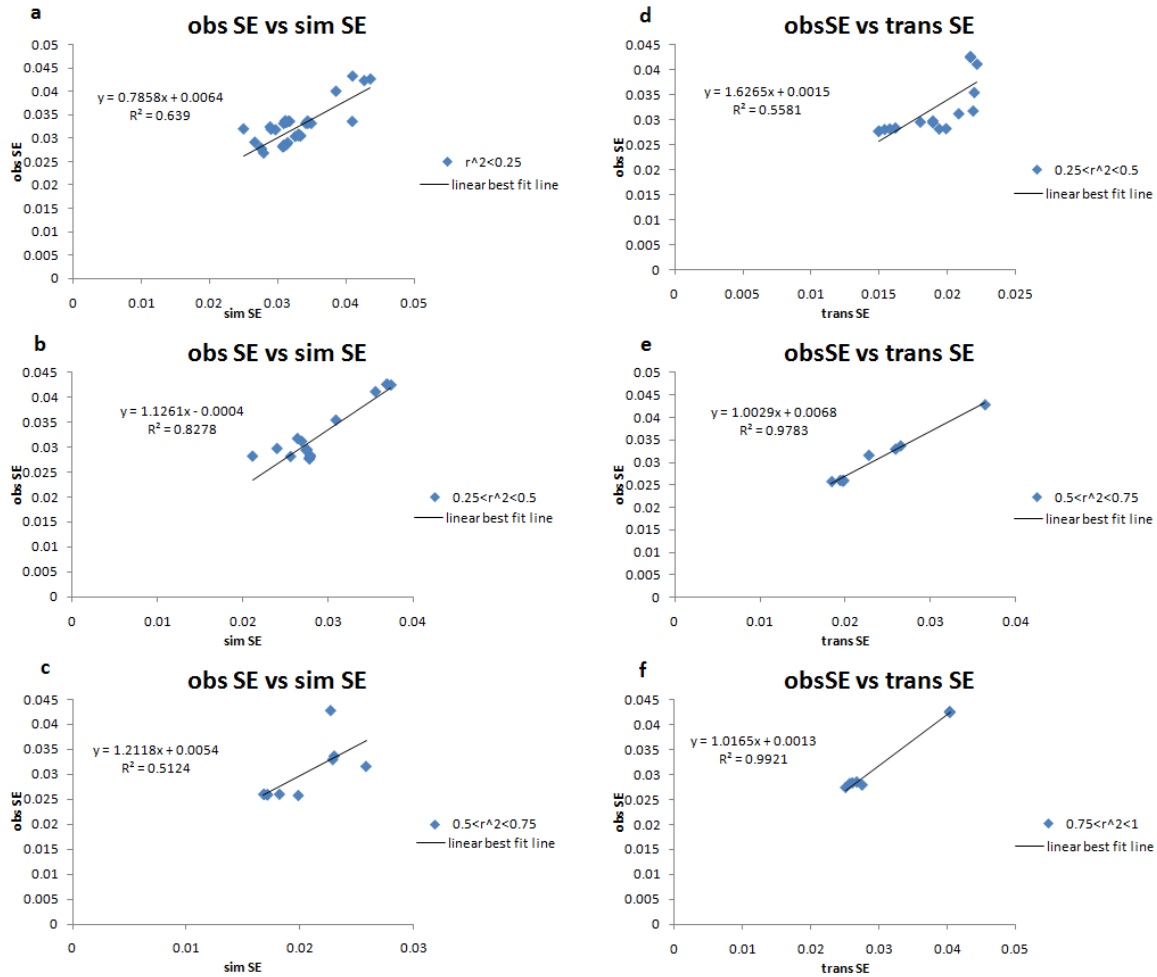


Figure 2-8. Comparison of the observed standard errors to the simulated standard errors and the transformed standard errors in different r^2 ranges

2.3.2. Relationship between r^2 and beta estimation

I illustrate the relationship between r^2 and the two beta estimations in the r^2 simulation I mentioned in Methods section. As shown in Figure 2-9, beta estimation value varies with different r^2 , and the shapes of these two beta estimations are similar when the observed

beta is smaller than 1.

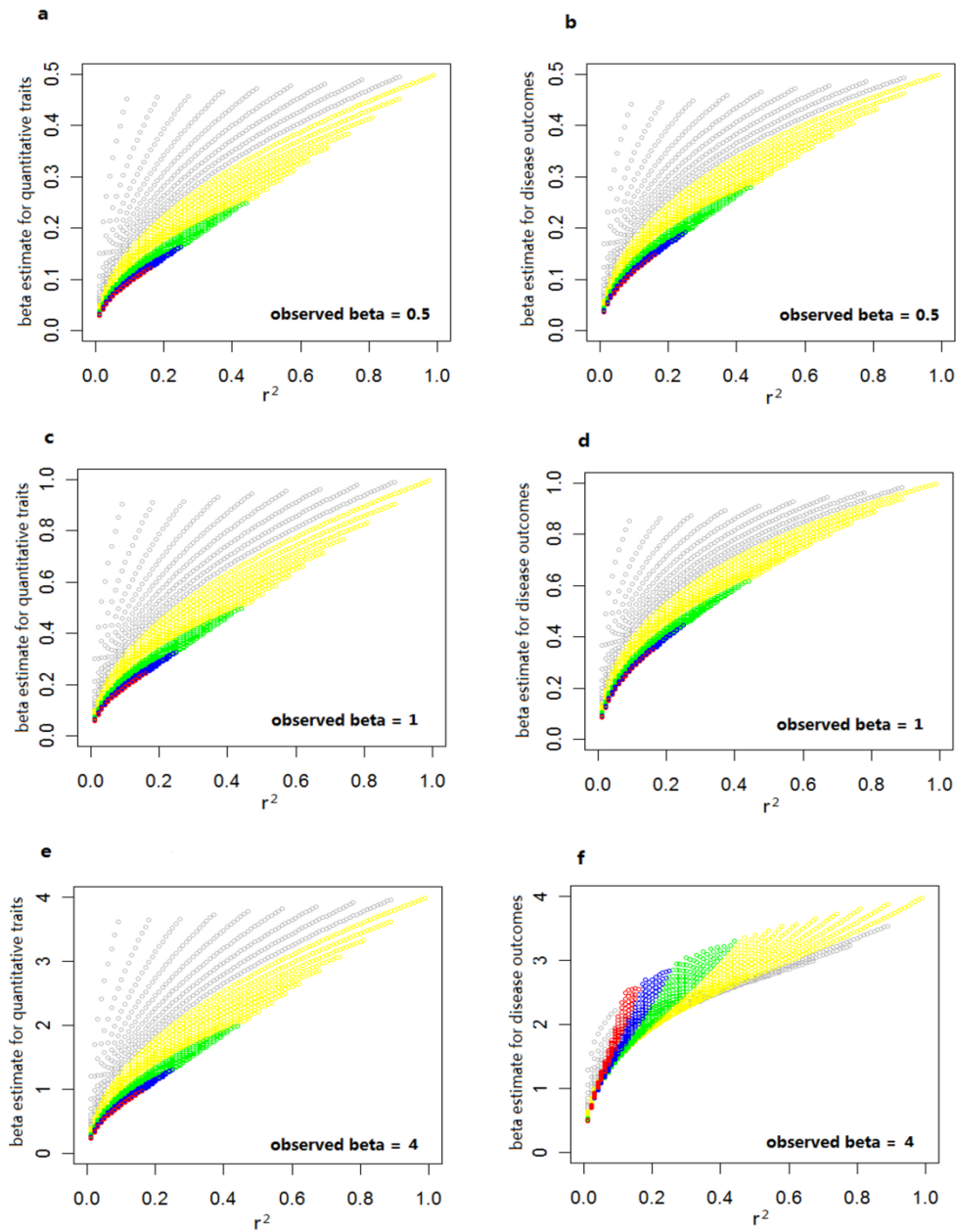


Figure 2-9. Relationship between r^2 and beta estimate of the possible dependent SNP for a given major allele frequency of the top hit and observed beta for the top SNP. Different colors refer to different intervals of major allele frequency of the possible dependent SNP (called q_1): red for $0.5 < q_1 \leq 0.6$, blue for $0.6 < q_1 \leq 0.7$, green for $0.7 < q_1 \leq 0.8$, yellow for $0.8 < q_1 \leq 0.9$ and grey for $0.9 < q_1 < 1$.

2.3.3. Example of a SSS-RAP plot

The SSS-RAP plot can visually show the relationship between each pair of SNPs. Figure 2-10 illustrated an example of the SSS-RAP plot. The top hit is the one on the right side with a bar of the observed beta only, whereas 6 test SNPs were in the left side. For each test SNP, the SSS-RAP plot compared the difference between the observed betas and the transformed betas. If two bars were overlap, this will suggest that the test SNP have some dependence conditional on the top hit. Otherwise, it suggests that the test SNP has certain level of independent effect conditional on the top hit. In this example, all test SNPs are suggested to be independent conditional on the top hit. However, the plot did not correct the multiple testing. In reality, we need to check whether the SSS-RAP p-value is smaller than the p value cut-off to conclude the independence of the test SNPs.

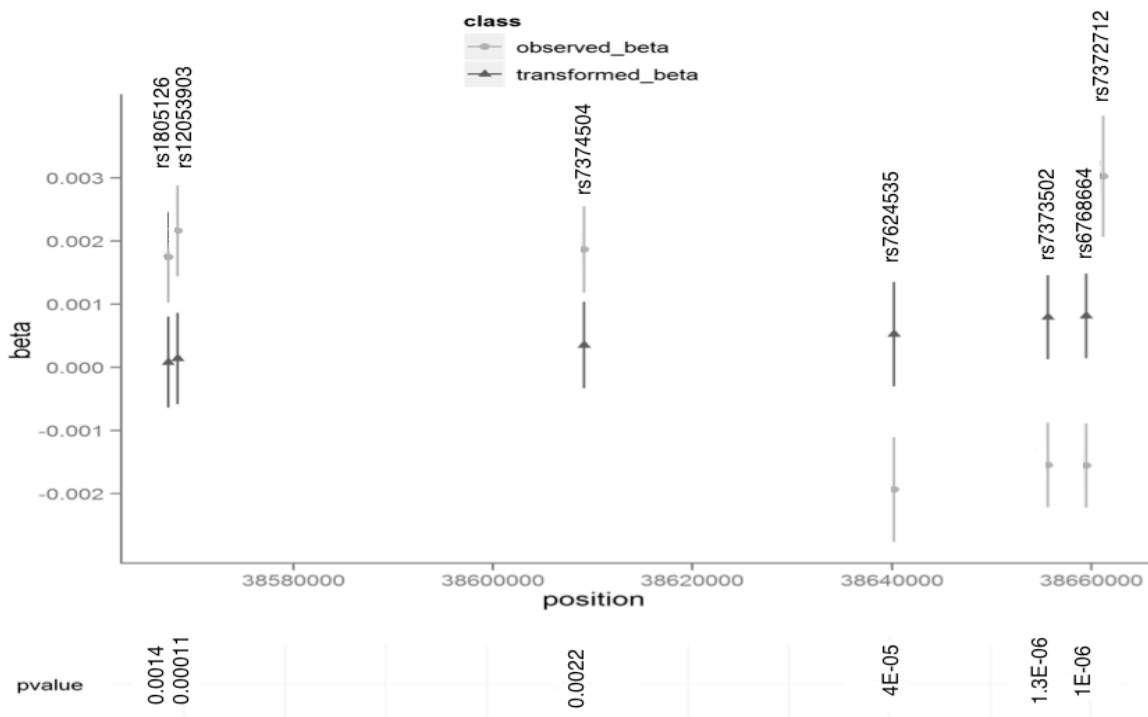


Figure 2-10. Example of a SSS-RAP plot

2.3.4. Performance comparisons of SSS-RAP, model selection methods and conditional analysis using BWHHS individual level data

The SSS-RAP program was used to analyse significant SNPs associated with lipid traits and ECG traits in BWHHS individual level data, which include SNPs in the lipoprotein lipase (*LPL*) (associated with both triglyceride and HDLc), cholesteryl ester transfer protein (*CETP*) (associated with HDLc), sodium channel, voltage-gated, type V, alpha subunit (*SCN5A*) (associated with PR interval and QRS duration) and nitric oxide synthase 1 (neuronal) adaptor protein (*NOS1AP*) associated with QTc interval (Table 2-4). Since the purpose of this analysis is a performance comparison rather than a discovery study (in which we need to be very careful about false positive), we selected a quite loose p value threshold so that we can include more SNPs into this analysis. Results from SSS-RAP were compared with ten different model selection criteria and conditional haplotype analysis.

genes	quantitative traits	mean	SD	N	significant SNPs	p-value threshold
<i>LPL</i>	TG (mmol/l)	1.873	1.251	3175	10	1×10^{-03}
<i>LPL</i>	HDLc (mmol/l)	1.660	0.453	3169	10	1×10^{-04}
<i>CETP</i>	HDLc (mmol/l)	1.658	0.455	3035	17	1×10^{-05}
<i>NOS1AP</i>	QTc interval (msec)	0.419	0.024	2686	24	1×10^{-04}
<i>SCN5A</i>	PR interval (msec)	0.159	0.023	2867	6	1×10^{-02}
<i>SCN5A</i>	QRS duration (msec)	0.091	0.015	2980	9	1×10^{-02}

Table 2-4. Summary information of BWHHS participants. P-value threshold is the cut off value of the SNP significant test. Abbreviations: mean, mean of quantitative traits; SD, standard deviation of quantitative traits; TG, triglycerides; HDL-C, high-density lipoprotein

cholesterol, $QTc = QT/\sqrt{RR \text{ interval}}$; N, participants after exclusion with missing genotypes and phenotypes.

BWHHS ECG traits:

For three ECG traits, I set the t test p-value of 0.05 as the elimination standard for SSS-RAP.

Generally, the SSS-RAP program selected the same optimal SNPs model as conditional analysis in all ECG traits. For *NOS1AP* loci associated with QTC interval and *SCN5A* SNPs associated with PR interval, the SNPs selected by SSS-RAP were also chosen by the other ten model selection methods (Figure 2-11 and Figure 2-12a). For *SCN5A* SNPs associated with QRS interval, BIC, CV, stepwise BIC and stepwise regression (Figure 2-12b) selected SNP rs10154914 instead of top hit rs7374540. This SNP is in some LD with the top hit ($r^2=0.153$).

BWHHS lipids traits:

For the three lipids traits, I set a t test p value of 0.01 as the threshold for backward elimination. As shown in Figure 2-12c, 3-11d and 3-11e, the results were consistent among the ten model selection methods, conditional analysis and SSS-RAP. The only exception was the top hit of *LPL* associated with TG, rs3735946. Stepwise AIC stepwise Cp, stepwise BIC and stepwise regression selected a coding SNP, rs328, instead (Figure 2-12c). Rs328 is in strong LD with top hit rs3735946 ($r^2=0.911$). Thus this finding is inconclusive in that this exception actually points to a 'real' independent signal.

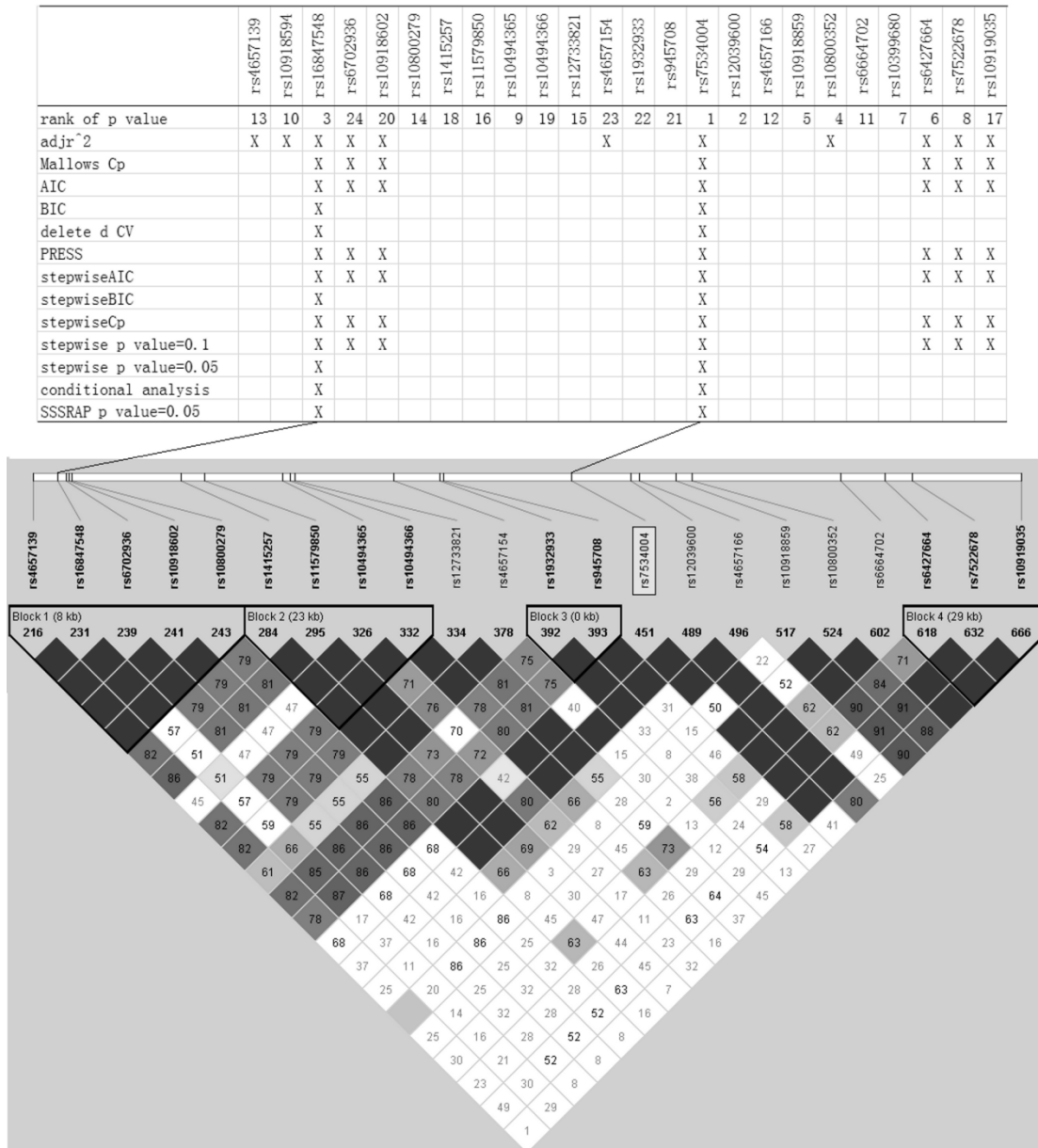


Figure 2-11. Comparison of model selection approaches, conditional analysis and SSS-RAP using BWHHS individual level data in *NOS1AP* region. All the methods and criteria used to select *NOS1AP* loci most likely to be tightly associated with QTc interval. For all the methods, cross (x) denotes inclusion in the best model. Top hit is represented by yellow background in LD plot. The P value cut-off of 0.05 was used in this analysis.

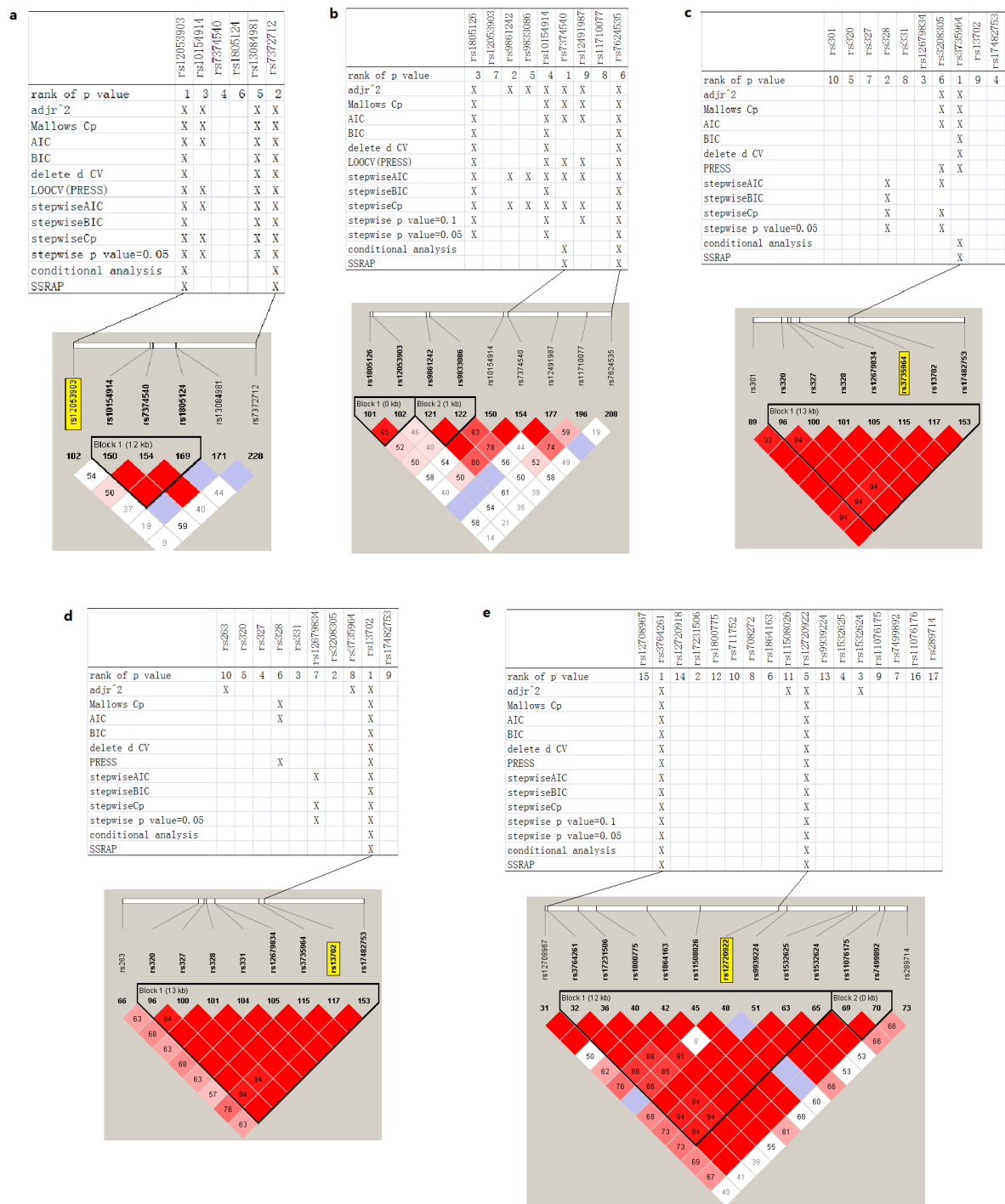


Figure 2-12. Comparison of model selection approaches, conditional analysis and SSS-RAP using BWHHS individual level data in other regions. All the methods and criteria used to select: **a**, *SCN5A* loci most likely to be independently associated with PR interval, the P value cut-off of 0.05 was used in this analysis; **b**, *SCN5A* loci most likely to be independently associated with QRS duration, the P value cut-off of 0.05 was used in this analysis; **c**, *LPL* loci

most likely to be independently associated with a TG variant, the P value cut-off of 0.01 was used in this analysis; **d**, *LPL* loci most likely to be independently associated with a HDL variant, the P value cut-off of 0.01 was used in this analysis; **e**, *CETP* loci most likely to be tightly associated with a HDL variant , the P value cut-off of 0.01 was used in this analysis. For all the methods, cross (×) denotes inclusion in the best model.

2.3.5. Applying SSS-RAP to an admixed population using 1000 Genomes Project data

For the analysis conducted on an admixed population using the 1000 Genome Project data, the principal components analysis (PCA) applied by the HapMap Consortium indicated that CEU and TSI populations are from a relatively homogeneous ancestry and can be grouped together under larger group of ‘European ancestry’ (International HapMap 3 Consortium et al., 2010).

In both populations SSS-RAP selected the same sets of independent SNPs in *SCN5A* for both PR interval and QRS duration (Table 2-5). Therefore, the minor departures in LD did not have a major effect to the SSS-RAP results.

	Step-wise stage	CEU only	CEU+TSI
<i>SCN5A</i> - PR	step 1	rs7372712	rs7372712
	step 2	rs7374540	rs7374540
	step 3	rs7624535	rs7624535
	step 4	rs12053903	rs12053903
<i>SCN5A</i> - QRS	step 1	rs7374540	rs7374540
	step 2	rs6797133	rs6797133

step 3	rs1805126	rs1805126
step 4	rs7624535	rs7624535

Table 2-5. *Test of one ancestral group assumption for SSS-RAP.* The SSS-RAP was applied using meta-analysis summary statistics of SCN5A associated with two traits. The reference panel was either HapMap CEU unrelated individuals or HapMap CEU plus TSI individuals. The SNPs in Step 1 to Step 4 refer to the independent SNPs selected by step-wise procedure 1 to 4.

2.3.6. Applying SSS-RAP to 210 simulated populations

The performance of SSS-RAP and conditional analysis were compared using 210 admixed populations using simulations (Chapter 3.2.3). For the three SNPs in this simulation, the top hit rs328 and the secondary independent signal rs327 was identified as independent by both method in all 201 populations. Interestingly, the performance of eliminating bystander SNP, were different for these two methods. In this simulation, the bystander SNP, rs263, was eliminated in 68.6% (144/210) of the simulations when applying the SSS-RAP algorithm (Figure 2-13a). In contrast, the conditional analysis did not eliminate the bystander SNP from any of the admixture populations (Figure 2-13b).

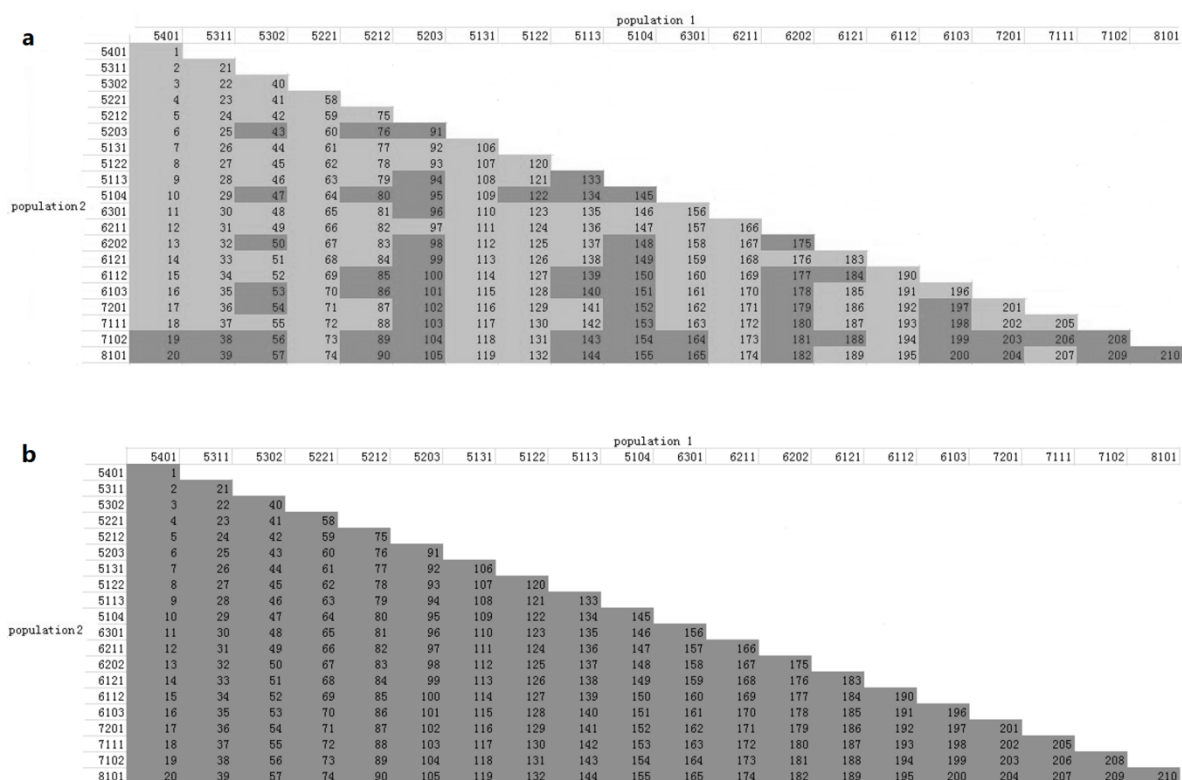


Figure 2-13. SSS-RAP and conditional analysis results for the three SNPs haplotype simulation.

a, represents the result for SSS-RAP. **b**, represents the result for conditional analysis. Since both methods working well for identifying the independent signals, so I only shown the difference in eliminating bystander SNPs in this plot. For each cell, light grey means the bystander SNP was eliminated by the method, which means the method return a result fit our setting perfectly. Dark grey the bystander SNP was not eliminated by the method, which means the method fail to achieve its function.

2.4. Discussion

In this chapter, I have utilized the basic concepts of LD measure (r^2) of pair-wise SNPs and regression coefficient (beta) to solve the problem of detecting independent SNPs conditional on the top hit using summary statistics from meta-analysis and reference genotype panel.

2.4.1. Advantages of SSS-RAP

Scope of SSS-RAP

Before the SSS-RAP program was developed, the researchers had to perform the conditional analysis at the level of individual studies; and then summary statistics from each of the individual-level studies are combined together through a second round of meta-analysis (Ma et al., 2010). Meta-analysis nowadays combines more and more single GWASs together, which makes the abovementioned process administratively laborious and time consuming (Zheng et al., 2013). The SSS-RAP algorithm extends the scope of fine mapping to group-level and enabled determining multiple associated variants using meta-analysis results directly.

Accuracy and inherent automation of SSS-RAP

Comparison between SSS-RAP and model selection methods

In the performance comparison using BWHHS individual-level data, all SNPs selected by SSS-RAP were chosen by most of the established model selection methods (Smith et al., 2010a) (Figure 2-11 and Figure 2-12). Only two top hits which were classified as ‘independent’ by SSSRAP were not replicated by the model selection methods. These methods chose SNPs in LD with the top hit instead. These differences can be attributed to the downsides of the other methods and not to SSS-RAP, since, (i) the results of SSS-RAP and conditional analysis were highly consistent using the individual-level data and (ii) SSS-RAP considers the biological correlation-LD, whereas model selection methods only consider the statistical correlation of the SNP models.

Comparison between SSS-RAP and conditional analysis

Among all the SNPs which were tested using BWHHS individual-level data and meta-analysis summary datasets (Figure 2-11 and Figure 2-12), 95.44% (293/307) of SSS-RAP results were consistent with conditional analysis results. This proves the highly accurate nature of SSS-RAP. Further analysing the inconsistent results, only two independent SNPs (rs1805126 and rs11710077) selected by conditional analysis which were not included in the SSS-RAP results. For SCN5A associated with PR interval, the coding SNP rs1805126 (D1819) was statistical significant conditional on the top hit rs7372712. But it is in strong LD with the second independent signal, rs12053903 ($r^2=0.827$). Similarly, rs11710077 is in modest LD with the second independent effect SNP rs6797133 ($r^2=0.386$) associated with QRS duration. As abovementioned, conditional analysis is administratively burdensome and time consuming. For the ECG meta-analysis, only the association analysis conditional on the top associated SNP at the loci were processed. So these two SNPs were still included in the model. In contrast, SSS-RAP is a step-wise elimination method. Like other step-wise regression strategy (Knüppel et al., 2012), it not only considers the independence conditional on the top SNP, but also considers the correlation between each possible independent effect SNP; and consequently excluded these two SNPs, which highlighted a major advantage over previous methods is the inherent automation in SSS-RAP.

Population structure causes genome wide LD between unlinked loci, which will cause statistical confounding in GWAS based fine mapping (Segura et al., 2012), so an additional performance comparison was conducted using admixture populations. Both SSS-RAP and conditional analysis shows the ability to detect the secondary signal in an admixture

population. Moreover, SSS-RAP shows the ability to eliminate bystander SNPs where conditional analysis failed to eliminating the bystander SNPs in any case (Figure 2-13). So I suggested that SSS-RAP is a more suitable fine mapping approach for admixture population analyses. This comparison illustrates the potential value of SSS-RAP when conducting fine mapping analysis using admixture populations who belong to the same ancestral group (e.g. European).

Computational time of SSS-RAP

As discussed above, conditional analysis using the conventional meta-analysis system is time consuming, which will normally take months for each round. Compared to conditional analysis, SSS-RAP is simple to use and yields results speedily. For SSS-RAP, there are no computationally unfeasible and/or expensive processes such as iterations and simulations thus the analysis usually takes a few seconds for each genomic region. Table 2-6 shows some examples of the number of SNPs in each model and the computational time of each analysis (the analysis result include the conducting of result table and the SSS-RAP plot) using a computer with 2.53GHz CPU and 2 GB RAM.

Loci	traits	number of SNPs	computational time (Second)
<i>SCN5A</i>	PR	6	2.54
<i>SCN5A</i>	QRS	10	3.1
<i>ABCG5-8</i>	GBD	10	3.99
<i>NOS1AP</i>	QTc	60	4.26

Table 2-6. *Computational time of SSS-RAP*

The web-based tool and visualization

As a user friendly program, SSS-RAP provides a lot of options for different user requests. Not

only can the users carry out analyses by using the SSS-RAP source code written in R, but also can use the web-based interface to analyze regions with less than 20 SNPs. As illustrated in Figure 2-10, SSS-RAP also provides a plot which shows the difference between expected beta and observed beta in the context of a regional association plot as well as the t test p-value. The plot shows the dependence effect of each test SNP “borrowed” from the top hit and helps the users to pick out the statistical independent effect SNPs easily.

2.4.2. Assumptions and limitations of SSS-RAP

One ancestry assumption

Different racial and ethnic and ancestral categories of human show differences arise directly through differing allele frequencies, which will influence LD patterns and haplotype frequencies (Race, Ethnicity, and Genetics Working Group, 2005). So the major assumption of SSS-RAP program is that participants of both the meta-analysis and the reference genotype panel are from one major ancestral group. Although I have proven that minor departure of LD will not affect the result of SSS-RAP significantly in section 3.3.5, users need to avoid using meta-analysis results from trans-ethnic analysis.

Assumption of the top hit

When applying SSS-RAP, I assume that the most significant SNP inside a genomic region represents a tag of a single causal signal or is itself the causal site. Conditional analysis shares the same assumption of the top hit. To fit this assumption better, it would be necessary to impute the genotypes before progressing to the meta-analysis so that the possibility of detecting the causal signal or a suitable proxy signal will be higher.

Limiting conditions for SSS-RAP and conditional analysis

The results of the admixed population simulation suggested that SSS-RAP and conditional analysis performs quite well if the top hit is the only independent effect SNP in the genomic region. If there are two or more independent loci in the region, SSS-RAP will eliminate most of the dependent SNPs. However, it is impossible to explain which loci are truly independent for the remaining SNPs. In fact, only haplotype-based analyses may be able to discriminate between causal sites, for example in the (most likely infrequent) situation where one “bystander” efficiently represents two separate causal sites themselves in LD (Figure 2-14).

Model 1

Effect score: $A=1$, $B=1$, $a=b=C=c=0$

causal SNP	causal SNP	bystander SNP	Haplotype effect
—A—	B—	C—	2
—A—	b—	C—	1
—a—	b—	c—	0
—a—	B—	C—	1

Model 2

Effect score: $C=4/3$, $A=a=B=b=c=0$

causal SNP	causal SNP	bystander SNP	Haplotype effect
—A—	B—	C—	4/3
—A—	b—	C—	4/3
—a—	b—	c—	4/3
—a—	B—	C—	4/3

Figure 2-14. An example of one ‘bystander’ efficiently represents two separate causal sites themselves in LD. In **Model 1**, bystander C shows the most signal. However, the haplotype effect scores are not consistent with C being causal because not all C bearing haplotypes show the same effect score. On the other hand, effect scores $A=B=1$ would be consistent with observed haplotype effects. In **Model 2**, C shows the most signal (same as in model 1). In this model, the haplotype effect score are consistent with C being causal, whereas A and B would not be.

2.4.3. Conclusion

In this chapter, I have successfully developed a novel fine mapping method, SSS-RAP, using meta-analysis summary statistics directly. I compared the performance of SSS-RAP to conditional analysis and model selection methods, most of the results are consistent among these methods, which proves the accuracy of SSS-RAP. A user-friendly web interface was developed by the same time. During the development of SSS-RAP, I also verified the linear relationship between MAF and the standard error of beta. Last but not least, the limitation of SSS-RAP made me realize that a haplotype-based fine mapping method was needed, which will be discussed in Chapter 5.

CHAPTER 3. FINE MAPPING OF CARDIOVASCULAR DISEASE RELATED TRAITS AND APPLICATION OF SSS-RAP

3.1. Introduction

Cardiovascular diseases (CVD), mainly coronary heart disease (CHD), stroke and rheumatic heart disease, are the leading cause of mortality and morbidity globally. The World Health Organization (WHO) reports that CVD deaths increased from 14.4 million in 1990 to 17.5 million in 2005. Among the CVD deaths, 7.6 million were caused by CHD and 5.7 million by stroke (WHO, 2009). The WHO further estimate that there will be about 20 million CVD deaths in 2015, accounting for 30% of deaths worldwide (WHO, 2005). CVD is the largest current contributor to global mortality and will continue increasing its percentage mortality in the future (WHO, 2009).

Finding the biological mechanisms of CVD and developing prevention strategies are of paramount importance.

Classic family and twin studies provide evidence that there is a heritable component contributing to CVD (Lloyd-Jones et al., 2004) (Marenberg et al., 1994) (Murabito et al., 2005) (Bachmann et al., 2012). In contrast to the above candidate gene approach analyses, GWAS provides a hypothesis free method of discovering novel molecular mechanisms of CVDs.

Previous GWAS studies investigating the impact of CVD looked at a range of different traits including Electrocardiography, body mass index and gallbladder disease. To contribute to this

cut-edge research field, I shall briefly introduce the background of these traits and the relative GWAS studies for these traits in the following section.

3.1.1. Electrocardiography (ECG)

Human heart beats regularly over a period of time. ECG is a quantitative measure of the electrical activity of the heart beating. It is measured by electrodes attached to the surface of the skin and recorded or displayed by a monitoring device external to the body (Cooper, 1986).

As shown in Figure 3-1, ECG traces the cardiac cycle, the quantitative measurement of ECG includes PR interval, QRS axis, QRS duration and QT interval (Hurst, 1998).

Electrocardiographic traits are important in predicting CVD morbidity and mortality (Lawlor et al., 2003) (Gaunt et al., 2012). A prolongation of QTc interval or QRS duration are linked to heritable determinants of risk of arrhythmias and sudden death (Mann, 2008). The mechanism of the common genetic variants contribute to ECG traits are unclear. Recent GWASs and meta-analyses successfully identified several genomic regions associated with ECG traits (www.genome.gov/gwastudies) although the causality unusually remain unknown.

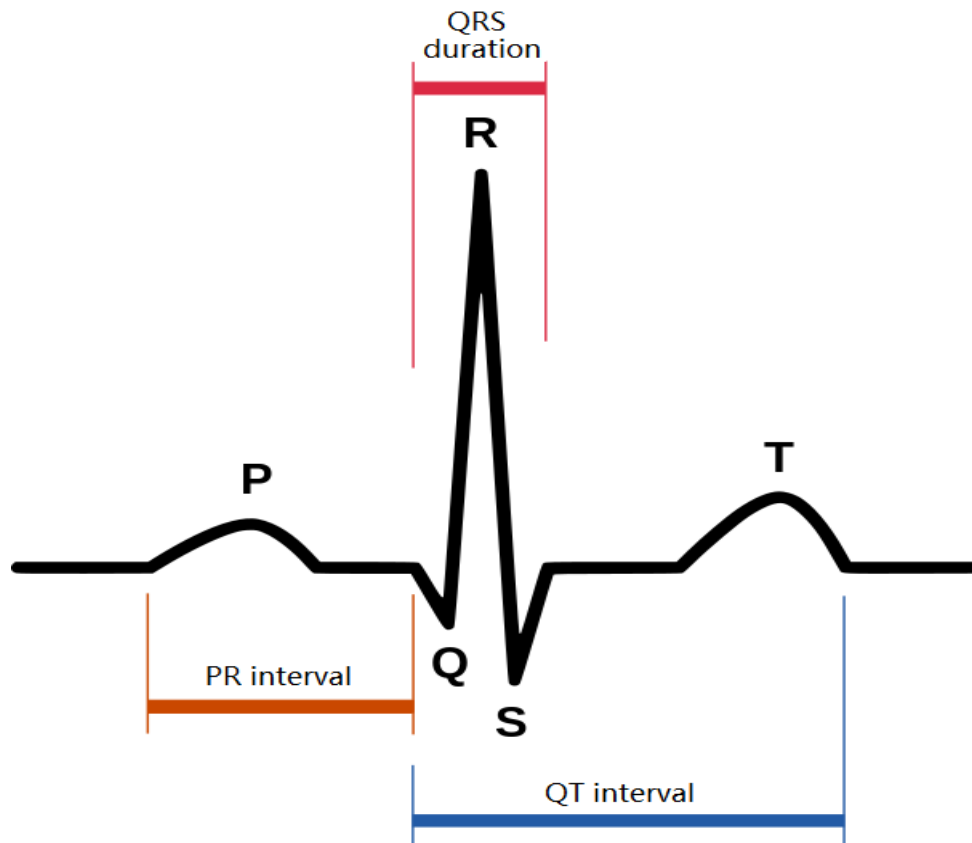


Figure 3-1. ECG waves and intervals.

PR interval

The PR interval is a measurement of the period from the beginning of the P wave to the beginning of the QRS duration. The PR interval represents ventricular depolarization. 27 SNPs were previously reported to be associated with PR interval (Pfeufer et al., 2010) (Smith et al., 2011) (Butler et al., 2012) (Jeff et al., 2013). As illustrated in Table 8-2, SNPs in *SCN5A* and *SCN10A* region got the strongest signals. *SCN10A* gene encodes the voltage-gated sodium channel Nav 1.8, which is essential for pain at low temperatures (Zimmermann et al., 2007). Moreover, the neighboring *SCN5A* gene encodes the main cardiac sodium channel, Nav 1.5. The mutations in *SCN5A* is related to Brugada syndrome, long-QT syndrome, dilated cardiomyopathy, cardiac conduction disease, idiopathic ventricular fibrillation and atrial

fibrillation (Remme et al., 2008). Moreover, Pfeufer A et al reported that there are more than one signal associated with PR interval within this region (Pfeufer et al., 2010).

QRS duration

QRS duration, also called QRS complex, reflects the rapid depolarization of the right and left ventricles. The ventricles have a large muscle mass compared to the atria, so the QRS duration usually has much larger amplitude than the P-wave. As listed in Table 8-3, 9 SNPs were previously reported to be associated with QRS duration (Jeff et al., 2013). In addition, Gaunt TR et al claimed that SNP rs7374540 in SCN5A is associated with QRS duration as well (Gaunt et al., 2012).

QT interval

The QT interval is measured from the beginning of the QRS complex to the end of the T wave. It represents ventricular repolarization and varies with heart rate and, for clinical relevance, requires a correction for this, giving the QTc.

$$QTc = \frac{QT}{\sqrt{RR}}$$

RR is the interval from the onset of one QRS duration to the onset of the next QRS duration, measured in seconds, often derived from the heart rate (HR):

$$RR = \frac{60}{HR}$$

Totally 77 SNPs are associated with QT interval (Arking et al., 2006) (Pfeufer et al., 2009) (Newton-Cheh et al., 2009) (Nolte et al., 2009) (Marroni et al., 2009) (Kim et al., 2012) (Smith et al., 2012) (Jeff et al., 2013). As shown in Table 8-4 A and B, SNP rs12143842, which is 6kb 5' of *NOS1AP*, got the strongest association with QT interval. *NOS1AP* (*CAPON*) is the

C-terminal PDZ domain ligand to neuronal nitric oxide synthase and it affects NMDA receptor-gated calcium influx (Jaffrey et al., 1998). *NOS1AP* expression in human left ventricular heart tissue (Arking et al., 2006). Common genetic variants inside *NOS1AP* are linked to sudden cardiac death in European population (Kao et al., 2009). Moreover, Newton-Cheh C et al claimed that, except top hit rs1214382, there are two additional independent SNPs associated with QT interval, rs12029454 and rs16857031.

3.1.2. Body mass index (BMI)

The BMI is a well-known index of measuring human obesity. It is defined as the body mass of a person divided by the square of his/her height. Its unit is kg/m².

$$\text{BMI} = \frac{\text{mass}}{(\text{height})^2}$$

As represented in Figure 3-2, the BMI provides a simple numeric measure of people's level of adiposity. A BMI less than 18.5 may indicate underweight. The BMI range between 18.5 and 25 are optimal weight. A BMI over 25 and under 30 is classified as overweight, whereas the BMI over 30 suggests the person is obese.

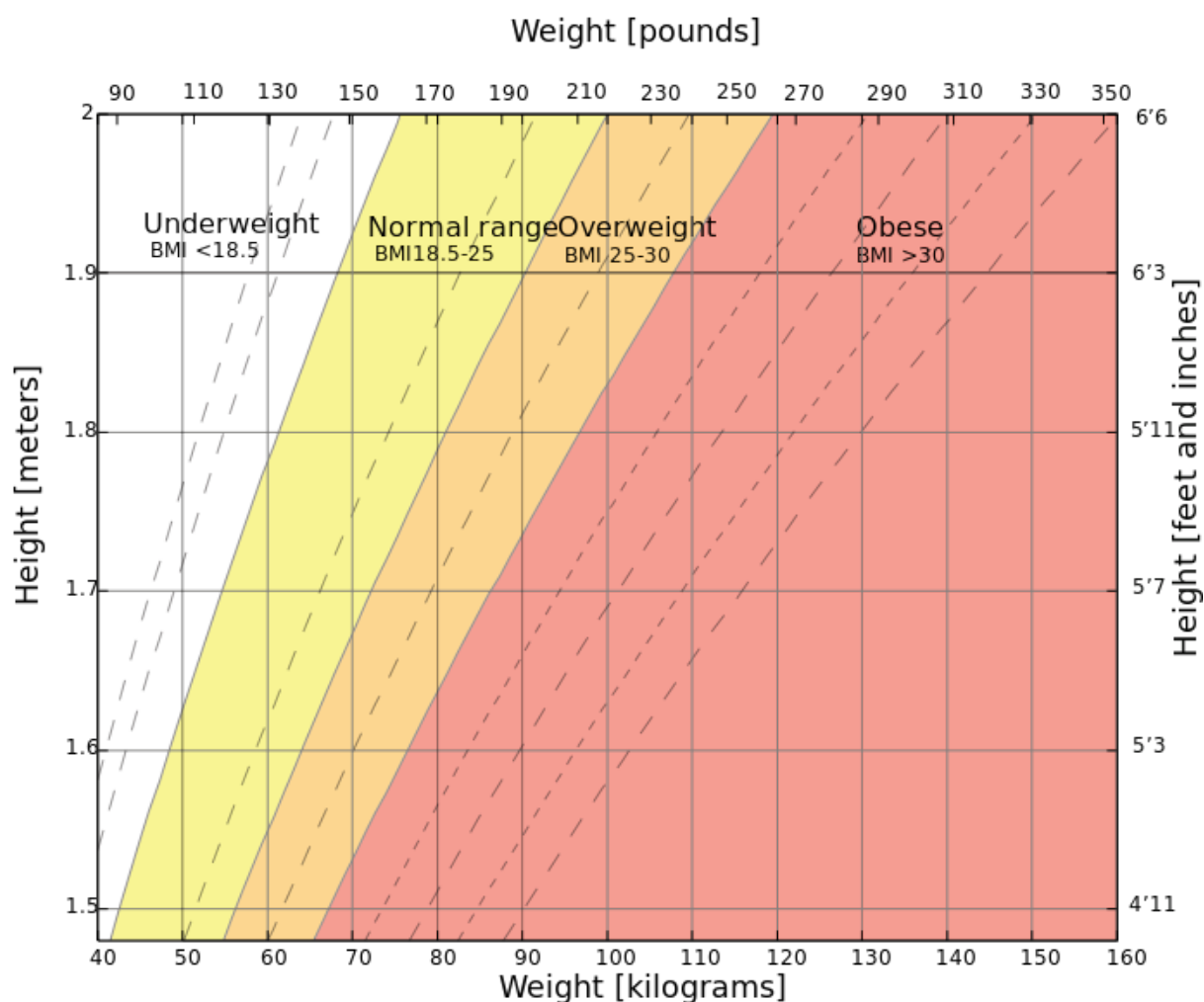


Figure 3-2. Body mass index chart. The scope of underweight, normal, overweight and obese are shown (http://en.wikipedia.org/wiki/File:Body_mass_index_chart.svg).

Obesity is one of the leading cause of preventable death in industrialised world (Richardson et al., 2013). The prevalence of obesity is increasing worldwide. In 2003 to 2004, the percentage of overweight and obese population in US was 66% and 32% respectively (Ogden et al., 2006). It is associated with an increased risk of diseases such as type 2 diabetes, heart disease, metabolic syndrome, hypertension, storks and so on (Frayling et al., 2007). Understanding its genetic influences will enhance insight into molecular pathogenesis. 19 studies identified approximately 150 SNPs associated with BMI. As listed in Table 3-1,

multiple SNPs near Fat mass and obesity associated (*FTO*) and melanocortin 4 receptor (*MC4R*) gene were reported to be associated with BMI (Thorleifsson et al., 2009).

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
16	53769677	FTO	rs6499640	A	5.25	4.00E-13
16	53816275	FTO	rs8050136	A	8.04	1.00E-47
18	57884750	MC4R	rs12970134	A	4.38	1.00E-12

Table 3-1. *FTO* and *MC4R* SNPs reported to be associated with BMI in the GIANT meta-analysis

3.1.3. Gallbladder disease (GBD)

The prevalence of gallstones is 10% - 40% in developed countries (Johnston and Kaplan, 1993). Gallstones lead to gallbladder disease (GBD), a major cause of morbidity, hospital admission, surgical intervention and economic burden. Causes of morbidity include biliary colic, cholecystitis, choledocholithiasis and pancreatitis (Kaloo and Kantsevov, 2001). In developed countries, most gallstones are formed of cholesterol. In contrast, pigment stones can be found in regions where bacterial and parasitic infections of the biliary tree, and hemolysis occur (Wang and Afdhal, 2004). Therefore, it will be interesting to see the relationship between lipids traits loci and GBD loci. Genome-wide association studies identified coding variant rs11887534 (D19H) in *ABCG5* and *ABCG8* region as the leading SNP affecting GBD (Buch et al., 2007). The head-to-head gene pair *ABCG5* and *ABCG8* together encode a heterodimeric transporter responsible for apical cholesterol secretion from both

hepatocytes and enterocytes. Rodriguez et al suggested multiple independent signals associated with GBD in this region (Rodriguez et al., 2014).

3.1.4. Aims and objectives

In chapter 3, I developed SSS-RAP and determined its accuracy using the performance comparison of conditional analysis, model selection methods and SSS-RAP. SSS-RAP is design for group-level data, so I shall focus on its application in group level in this chapter. Before the real case applications, I shall compare the SSS-RAP meta-analysis results to its individual-level results using different data sets as reference panel. For the CVD related traits I introduced above, SSS-RAP shall be applied to the candidate regions that have been associated with multiple signals by previous GWASs and meta-analysis, such as *NOS1AP* and *SCN5A* associated with ECG traits, *FTO*, *MC4R* and other regions associated with BMI as well as *ABCG5* and *ABCG8* region associated with GBD.

Objectives

1. Verification of the performance of SSS-RAP in meta-analysis level using summary statistics of ECG traits and GBD outcomes. Compare the meta-analysis level results with its individual-level results using different data sets as reference panel, for example, data from a specific study and publically available data
2. Discover the relationship between correlation coefficient, effect size and p value using GBD meta-analysis data
3. Fine mapping the genetic variants in *NOS1AP* and *SCN5A* regions associated with ECG condition traits using meta-analysis level data

4. Developing a transformation method so that the GIANT BMI meta-analysis summary statistics can be applied to SSS-RAP
5. Detecting LD pattern of SNPs significantly associated with GBD in the meta-analysis and identifying independent signals of GBD outcomes in *ABCG5* and *ABCG8* region using SSS-RAP

3.2. Methods and Materials

3.2.1. Sample datasets

The BWHHS individual-level genotypes

Among all the BWHHS individuals, I used genotype data of 3445 individuals who were genotyped using the Human cardiovascular disease (HumanCVD) BeadChip (Illumina, San Diego, CA). MAF were calculated using PLINK option --freq and haplotype frequencies were calculated using MIDAS (Gaunt et al., 2006).

Meta-analysis of ECG traits

The summary statistics of ECG meta-analysis were used (Gaunt et al., 2012). The analysis results of PR interval, QRS duration and QTc interval were used here. All the significant SNPs for this meta-analysis were listed in Table 8-5.

GIANT BMI meta-analysis summary statistics

The GIANT consortium released the summary data for their meta-analysis of GWASs. For BMI meta-analysis, this includes approximately 2.8 million SNPs in up to 123856 individuals in discovery panel and additional 125931 individuals in the follow-up study. The summary

results of this meta-analysis were downloaded from the consortium webpage (http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files). The summary results were consisted of the following information for each SNP: marker ID, allele 1, allele2, allele frequency of Allele 1, p values and number of observations. The standard errors of the betas were not included in the summary results so were calculated analytically.

Meta-analysis of GBD outcomes

The meta-analysis of GBD genotyped 15213 women of European ancestry (3216 cases and 11997controls) using the Human cardiovascular disease (HumanCVD) BeadChip (Illumina, San Diego, CA), which containing up to ~53000SNPs. The odds ratios with p values for development of GBD were generated (Rodriguez et al., 2014).

3.2.2. Applying SSS-RAP to meta-analysis of ECG traits and performance comparison of SSS-RAP in group level

Fine mapping of ECG traits

The SSS-RAP program was applied to the summary statistics of the meta-analysis on the ECG traits (Gaunt et al., 2012). I selected three quantitative ECG traits (PR interval, QRS duration and QTc interval) and decide to fine map three gene regions containing SNPs associated ($p < 5 \times 10^{-4}$) with these traits.

Performance comparison of SSS-RAP using meta-analysis data and different reference panels

Utilizing these meta-analysis data, I generated four data-sets using different combinations of: BWHHS individual level data (as the “in-house” component of the meta-analysis, individual level data were available), meta-analysis summary data (i.e. group-level betas and standard errors) and HapMap LD information (haplotype frequency X_{11} from HapMap individual level data). The compositions of these data-sets are described in Table 3-2.

Number of data-set	beta and SE	MAF	Haplotype freq (X_{11})
1	BWHHS individual level data	BWHHS individual level data	BWHHS individual level data
2	Meta-analysis	BWHHS individual level data	BWHHS individual level data
3	Meta-analysis	BWHHS individual level data	HapMap
4	Meta-analysis	HapMap	HapMap

Table 3-2. Content of combination of different data-sets, X_{11} represents the haplotype frequencies of two major alleles of pair-wise SNPs. Meta-analysis means data selected from meta-analysis summary data. Internal data means a single study was involved in the meta-analysis. HapMap means data calculated using HapMap CEU individual level data.

3.2.3. Applying SSS-RAP to GIANT meta-analysis of BMI

To utilise the summary statistics of GIANT meta-analysis of BMI, I need to find a way to simulate the betas and the standard errors of betas, which is not listed in the original summary results.

Simulation of the standard errors

As explained in Chapter 3.3.1, there is a clear positive correlation between minor allele frequency and standard error of beta (Figure 2-7). Fortunately, the standard errors of betas of 32 significant SNPs were found from the BMI meta-analysis paper (Speliotes et al., 2010). Using these data, I was able to simulate a relationship between SE and MAF for these 32 SNPs (Figure 3-3). Identifying the general relationship between SE and MAF will be an interesting follow up topic.

As shown in Figure 3-3, a power function fitted the relationship between SE and MAF with the least error ($R^2 = 0.6486$). Since I only found SE for a relative small number of SNPs for the GIANT BMI dataset, I expected a relatively noisy relationship between MAF and SE. However, it is an interesting attempt for meta-analysis with missing data. Using this power function, I simulated the standard errors of betas for the 2,284,195 SNPs using the minor allele frequencies provided by the database.

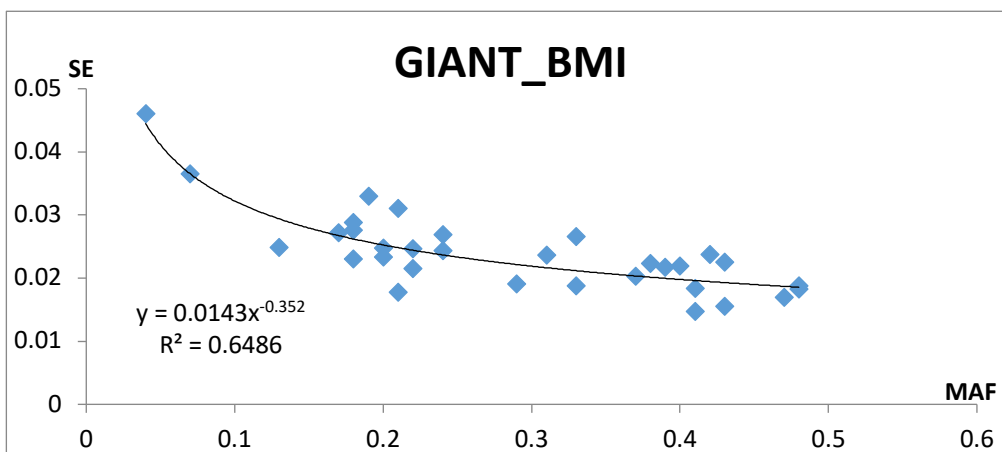


Figure 3-3. Predicting standard error using minor allele frequency. X-axis is the minor allele frequencies of obtained from the meta-analysis. Y-axis is the standard error. The relationship between MAF and SE are empirical and not perfect.

Beta prediction

The P-value provided in the GIANT BMI dataset was transformed from the t statistic. Here I use the reverse process to get the t statistic from the p-value (degree of freedom is N-2, where N is the number of individuals). Then I was able to simulate Beta directly using the existing data:

$$\text{beta} = \text{t statistics} \times \text{SE of beta}$$

Where the standard error of beta were calculated using the prediction formula in Figure 3-3.

The flow chat of the whole beta prediction process was illustrated in Figure 3-4.

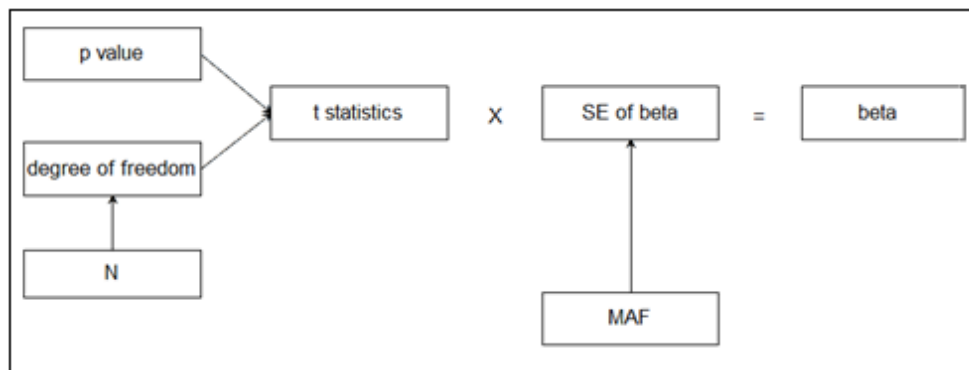


Figure 3-4. Flow chart of the beta prediction

Using the above process, I simulated Betas with their standard errors in order to run SSS-RAP (See discussion for details). In addition, I used the R package “NCBI2R” (<http://cran.r-project.org/web/packages/NCBI2R/index.html>) to estimate LD correlation (r^2) from HapMap and used Biomart web tool (<http://www.biomart.org/>) to find the coordinates for the SNPs. I then performed SSS-RAP on the remaining SNPs. T test p-value of 0.1 was set as the threshold of elimination for this dataset.

The SSS-RAP program was applied to the adjusted summary statistics of this meta-analysis. 32 significant SNPs for BMI were selected as 32 top hits. To take into account the possibility

of long distance LD, all SNPs inside the region of $\pm 5\text{Mb}$ (million base pair) away from each top hit were picked out for a further SNP independent test.

3.2.4. SSS-RAP analysis of GBD traits and identifying of relationship between GBD associated SNPs and lipids associated SNPs

Independent SNP identification

Within the ABCG5-8 region 360 SNPs were analyzed. After excluded SNPs with perfect LD ($r^2=1$), a total of 118 SNPs remained in the model. I analyzed the remaining SNPs using SSS-RAP. Using this approach 10 out of the 118 SNPs remained significant as signals in the meta-analysis (Table 8-6). Both HapMap and BWHHS individual level genotypes were used as reference panel for performance comparison.

Relationship of GBD associated SNPs and lipids associated SNPs

Since most gallstones are formed of cholesterol, I investigate the relationship between GBD and plasma lipids level in this section. From the 95 loci previously reported to be independently associated with blood lipid levels (28), it was possible to subject 63 to a test using IBC array data (Table 8-7).

The quantile-quantile (QQ) plots for GBD meta-analysis p value of these 63 were performed in R using the 'ggplot2' package (<http://cran.r-project.org/web/packages/ggplot2/index.html>).

To test whether the relationship between GBD and blood lipids traits was not by random

chance, I designed a simulation. The null hypothesis was that in QQ plots and compared with appropriate random sampling of equivalent numbers of SNPs in the IBC chip, the distribution of significance values would be no different for a trait-selected set of SNPs compared with a random set. However, considering the association of gallstones with obesity and their constitution of cholesterol, I hypothesized either a general difference of distribution, and/or that the leading SNPs for the trait would also show nominal significance for GBD.

For the simulation, I firstly ordered the 63 p-values ascendingly and then I calculated the observed p value and expected p value:

$$Obs_n = -\log_{10}(P_n)$$

$$Expt_n = -\log_{10}\left(\frac{n}{N}\right)$$

Where N is total the number of SNPs and n represent the current SNP inside 1 to N. Obs_n is the observed p value of SNP n, $Expt_n$ is the expected p value of SNP n.

Then the median number of the p values of 63 SNPs were calculated as follows:

$$\text{median} = \frac{\text{median of observed p value}}{\text{median of expected p value}}$$

In comparison with this median, I drew a distribution of median $\log_{10}p$ values generated by 10,000 simulations using random sampling of 63 loci for each set of simulations. These simulations were performed in R. Representations of the histogram of median estimates was performed using the 'hist' function in the basic package of R (<http://stat.ethz.ch/R-manual/R-devel/library/graphics/html/hist.html>).

3.3. Results

3.3.1. Applying SSS-RAP program to ECG meta-analysis summary results and four data combinations

After applying SSS-RAP to meta-analysis of ECG traits, I found that generally the results of SSS-RAP are highly consistent with conditional analysis results using group-level results.

***NOS1AP* associated with QTc interval**

For SNPs associated with QTc interval in *NOS1AP* region, conditional analysis of the meta-analysis selected 3 SNPs from 3 different LD blocks. In individual-level, SSS-RAP only detected two SNPs. In group-level, SSS-RAP return same SNPs model using meta-analysis results with BWHHS MAF and haplotype frequencies. When using HapMap MAF and haplotype frequencies instead of BWHHS ones, 3 more SNPs were detected, showing some uncertainty (Figure 3-5).

For SNPs associated with QRS duration in *SCN5A* region, first round conditional analysis selected 5 SNPs. Using meta-analysis summary statistics, SSS-RAP detect four of them: SNP rs11710077 was omitted by SSS-RAP, which have been explained in Chapter 3.4.1. In contrast, SSS-RAP only selected 2 SNPs using individual-level data (Figure 3-6 b).

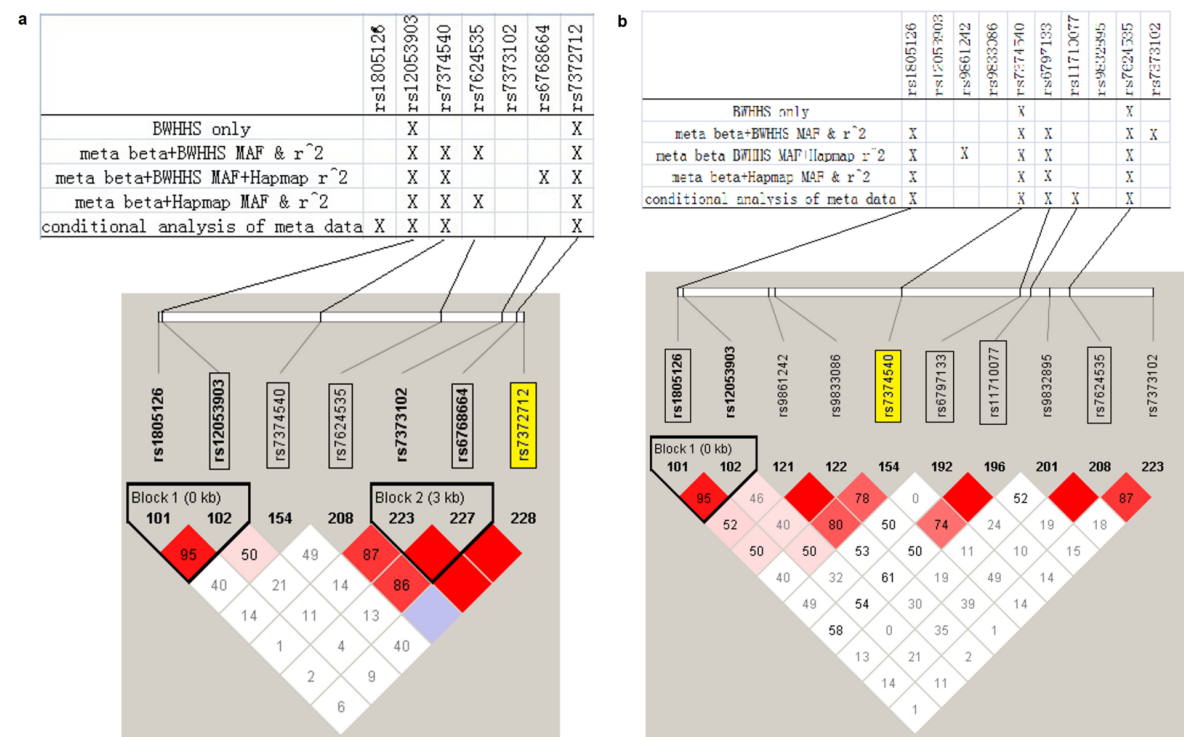


Figure 3-6. Comparison between SSSRAP and conditional analysis using meta-analysis data in ECG traits. Linkage disequilibrium and haplotype block structure were linked to related SNPs in **(a)** and **(b)**, *SCN5A* loci associated with PR interval and QTc interval respectively. Cross (x) denotes inclusion in the best model for each data combination.

3.3.2. SSS-RAP analysis of GIANT meta-analysis of BMI

As mentioned in Chapter 4.2.3, I tested the independence of SNPs near the 32 significant signals identified by the original meta-analysis paper. The results for the regions with

secondary signals are shown in Table 3-3. SSS-RAP identified 5 regions with more than one association signals inside each region.

SNP	NOTE	GENE	CHR	POS_37	REF_ALLELE	FREQ	SSS-RAP_P
rs887912	top hit	<i>FANCL</i>	2	59305625	T	0.325	0
rs12986742	independent snp1	<i>FANCL</i>	2	58975143	C	0.5	3.00E-08
rs11688816	independent snp2	<i>FANCL</i>	2	63053048	A	0.5417	2.54E-08
rs206936	top hit	<i>NUDT3</i>	6	34828553	T	0.7417	0
rs2033529	independent snp1	<i>NUDT3</i>	6	40348653	G	0.2583	8.78E-09
rs10767664	top hit	<i>BDNF</i>	11	27694241	G	0.825	0
rs1519480	independent snp1	<i>BDNF</i>	11	27675712	C	0.3667	3.52E-10
rs12444979	top hit	<i>GPRC5B</i>	16	19935389	A	0.1333	0
rs11074446	independent snp1	<i>GPRC5B</i>	16	20255123	C	0.175	1.14E-09
rs571312	top hit	<i>MC4R</i>	18	57838401	A	0.2833	0
rs17773471	independent snp1	<i>MC4R</i>	18	57977377	T	0.3333	4.39E-08
rs7227255	independent snp2	<i>MC4R</i>	18	58055731	G	0.9833	1.75E-10

Table 3-3. Genes that found extra independent effect SNPs by SSS-RAP program in GIANT BMI database.

Top hits are found from previous paper, 2nd-6th means the second to sixth independent effect SNPs.

3.3.3. SSS-RAP analysis of GBD associated SNPs and analysis of loci associated with lipids traits

LD pattern of SNPs significantly associated with GBD

For the ten SNPs significantly associated with GBD, I sketched the LD pattern among them and try to find out the relationship between correlation coefficient (r^2) and effect size (and r^2 vs p values as well).

The top associated SNP rs4953023 is in perfect LD ($r^2=1$) with the coding variant rs11887534 (representing an Aspartic acid to Histidine change at amino acid residue 19: D19H) that was associated with cholesterol gallbladder stones in previous GWAS and linkage studies (Grünhage et al., 2007) (Katsika et al., 2010). Rs4953023 is in intron 3 of *ABCG8*. It yields

both the highest significance ($P=3.57 \times 10^{-48}$) and the highest effect size ($\beta=0.763$) in the meta-analysis study. Figure 3-7A shows r^2 between each of the *ABCG5* and 8 SNPs significantly associated with GBD in meta-analysis and the top hit rs4953023. As shown in the figure, there are almost no LD between them. In addition, the effect sizes and p values of the other 9 SNPs are not comparable to those of the top hit. This observations support rs4953023 is likely independent of the other *ABCG8* SNPs that are significantly associated with GBD in the meta-analysis study.

One of the *ABCG8* SNPs, rs4299376, was found to be significantly associated with GBD ($P=3.10 \times 10^{-18}$) in our study with a lower effect size ($\beta=0.283$) observed. This SNP is in some LD with a coding SNP rs4148211 (BWHHS $r^2=0.209$). Rs4148211 (a Cysteine to Tyrosine change at residue 54: C54Y) is reported to be associated with GBD in a Taiwanese study (Kuo et al., 2008). Moreover, in HapMap CEU population, the observed LD between this pair of SNPs is $r^2=0.161$, $D'=1$. Our results support that these two SNPs may be in one LD block and may represent the same functional mechanism or (unknown) causal SNP.

Identifying the LD pattern step wisely, I removed the top hit from the SNPs model. Figure 3-7B shows LD between rs10439467 in intron 10 of *ABCG5*, and other *ABCG5/8* SNPs except top hit rs4953023. Rs10439467 appears to mark a novel independent effect that does not appear to have been directly represented or captured in earlier chip-based analyses. There is essentially no LD between this SNP with rs6720173 E604Q ($r^2=0.095$ in Europeans in HapMap), which is the only reported coding SNP for GBD in *ABCG5*. This confirmed its independence of effect on GBD. Besides, four other significant SNPs in the *ABCG5* region were no longer retained as independent predictors, respectively rs4953019 ($p=4.61 \times 10^{-11}$,

effect size=0.3406, $r^2=0.352$), rs10208987 ($p=4.77 \times 10^{-14}$, effect size=-0.3782, $r^2=0.352$), rs10201851 ($p=1.54 \times 10^{-09}$, effect size=-0.2629, $r^2=0.196$) and rs4148189 ($p=8.66 \times 10^{-09}$, effect size=0.2538, $r^2=0.196$).

Figure 3-7C shows LD between the remaining three SNPs with rs6720173 (E604Q) ($P=3.23 \times 10^{-11}$, $\beta = 0.256$). Two significant SNPs rs2278357 ($r^2=1$, $P=1.10 \times 10^{-9}$, $\beta = 0.233$) and rs4245786 ($r^2=0.234$, $P=1.88 \times 10^{-5}$, $\beta = -0.142$) are in LD with E604Q. The previous possible independent SNP rs10439467, is in very low LD with E604Q ($r^2 = 0.095$ in Europeans in HapMap).

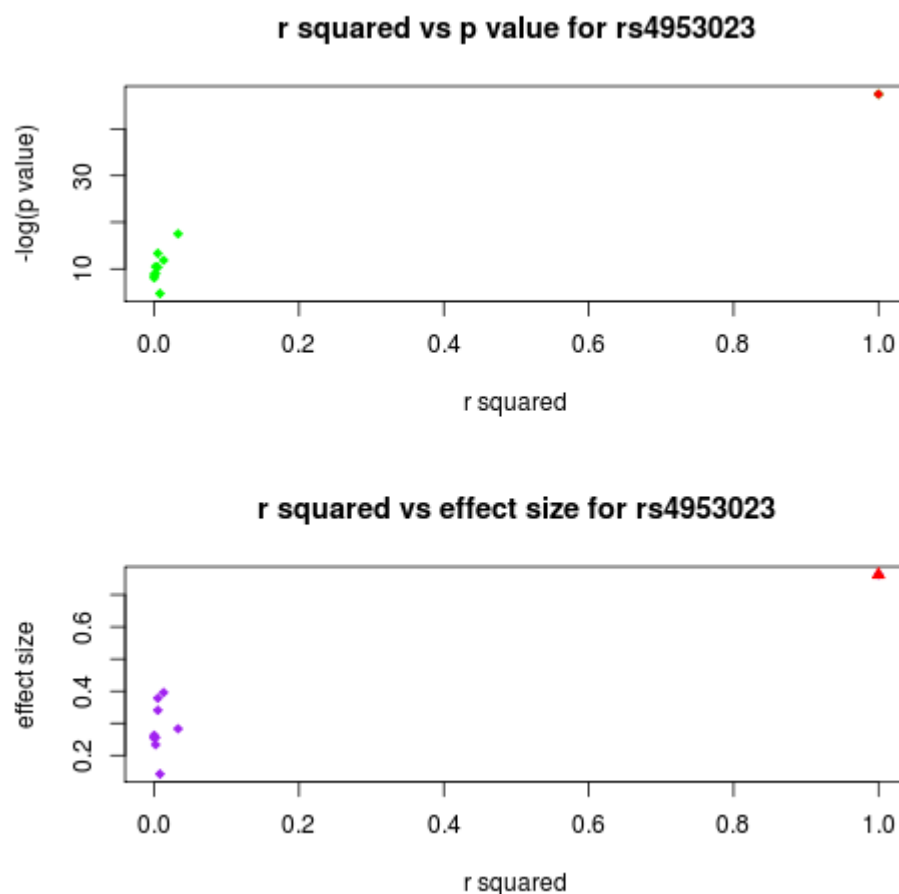
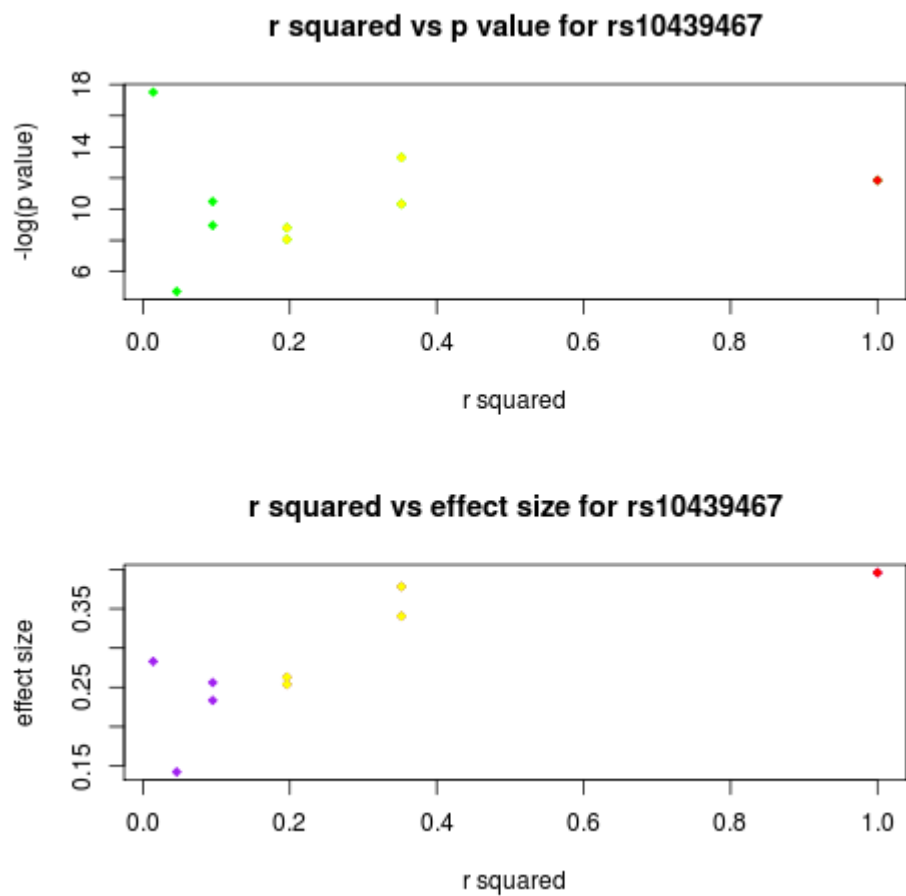
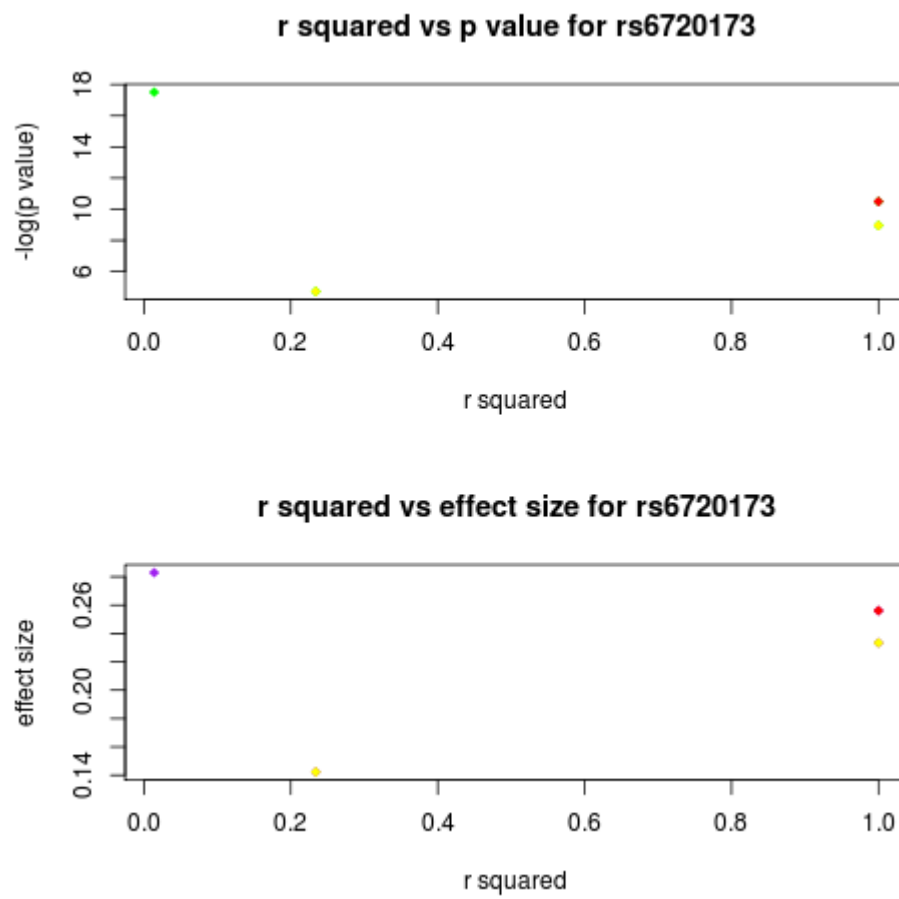


Figure 3-7. Comparison between r^2 values and P values and between r^2 values and effect sizes for key SNPs associated with GBD

(A) LD between the top hit rs4953023 and the *ABCG5* and 8 SNPs significantly associated with GBD in meta-analysis



(B) LD between rs10439467 in intron 10 of *ABCG5*, and other *ABCG5/8* SNPs except top hit rs4953023



(C) LD between coding SNP rs6720173 (E604Q) and other *ABCG5/8* SNPs

SSS-RAP analysis of SNPs associated with GBD

Here I compared the performance of SSS-RAP in individual-level and group-level using GBD outcomes. As shown in Table 3-4, when applying SSS-RAP to individual-level data, only two SNPs were detected. In addition, as shown in Chapter 4.3.3 LD pattern section, I found four SNPs, rs4953023, rs4299376, rs10439467, rs6720173, as possible independent effect variants. Similarly, SSS-RAP detected the same four SNPs as independent in *ABCG5* and *8* region when using BWHHS individual-level genotypes as reference panel (Table 3-4). Moreover, when using meta-analysis summary statistics and HapMap genotypes as reference panel, five SNPs were detected. The extra SNP rs4953019 is in modest LD (HapMap CEU $D' = 1$, $r^2 = 0.312$) with the coding SNP rs6720173. This result suggest that these two SNPs are in one LD block and may represent the same functional mechanism.

dataset	reference panel	phase 1	phase 2	phase 3	phase 4	phase 5
BWHHS	BWHHS	rs4953023	rs4299376			
meta-analysis	BWHHS	rs4953023	rs4299376	rs10439467	rs6720173	
meta-analysis	HapMap	rs4953023	rs6720173	rs10439467	rs4299376	rs4953019

Table 3-4. Comparison of performance of SSS-RAP in individual-level, meta-analysis level and different reference panels in GBD traits.

Detecting relationship between GBD and blood lipids

Figure 3-8A shows the QQ plots for the meta-analysis p value of 63 SNPs listed in Table 8-7. These SNPs represents 63 independent loci previously reported to associate with plasma lipid levels. The observed p values was significantly different than the expectations. Out of 63 SNPs, 59 showed an observed p value above the expected p value in a QQ plot. This is in accordance with a combined effect of these SNPs both on GBD and on plasma lipid levels

variation.

To test whether the significant of the 63 SNPs were not by random chance, I simulated a median distribution (Chapter 4.2.4) Figure 3-8B shows results for simulations analysing the median $\log_{10}p$ value of 10,000 samples of groups of 63 random SNPs in IBC. The 10000 random simulations got a distribution with mean of 1.07 (median =1.06) and standard deviation of 0.19. In contrast, the observed median value for the 63 lipid-associated loci was 1.494. A median p value greater than that observed for the lipid loci set, was only appeared in 1.7% of random samplings. So the histogram in Figure 3-8B conformed the significance of p values I observed by the QQ plot shown in Figure 3-8A.

Further linkage between GBD and blood lipids can be found in Table 8-7. The loci influencing either LDLc or triglycerides showed the most outstanding associations with GBD. Specifically, these were rs4299376 in ABCG5/ABCG8; rs2081687 in CYP7A1 (the proxy SNP in the meta-analysis rs8192870 has an $r^2 = 0.925$ with rs2081687, $P=0.0031$ for GBD); rs1260326 in GCKR ($P=0.00013$); rs5756931 in PLA2G6 (the proxy SNP in the meta-analysis rs4820314 $r^2 = 0.8$ $P=0.0048$ for GBD); rs2479409 in PCSK9 ($P=0.021$) and rs1532085 in LIPC ($P=0.036$).

This analysis is quite similar to a Mendelian Randomization (MR) approach, which is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in non-experimental studies (Davey Smith and Ebrahim, 2003). So there is a possibility to process a formal MR to further corroborate the causal gene. However, it is out of the scope of this work.

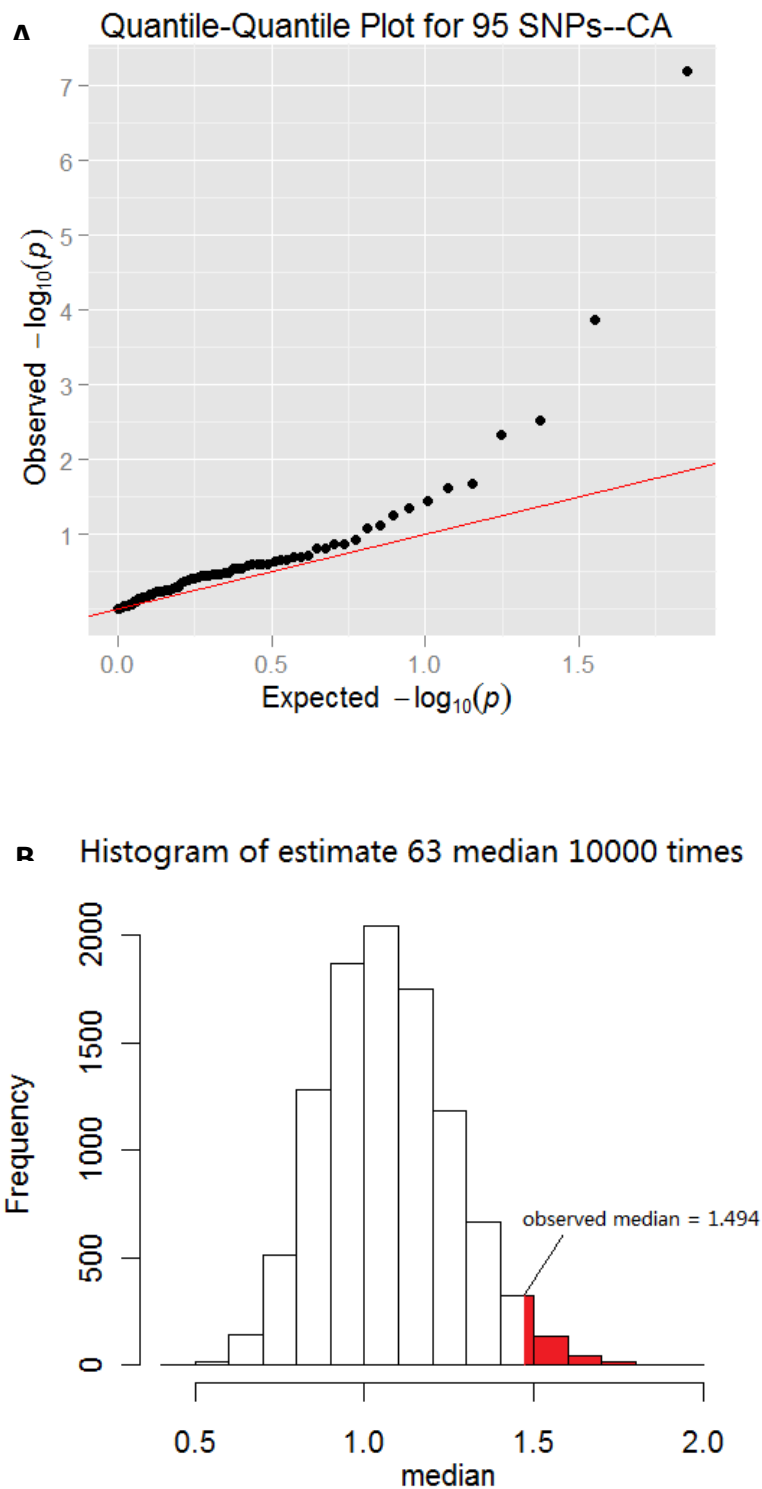


Figure 3-8. Pathway analysis relating GBD with SNPs associated with plasma lipid levels

(A). GBD QQ plot for SNPs representing 63 independent loci reported in a published study to influence plasma lipid levels.

(B). Simulation of the median \log_{10} p value when sampling 63 SNPs at random from the IBC array meta-analysis, 10,000 random samplings. The observed median was 1.49. The median observed for the lipid set was 1.06 (mean = 1.07, SD = 0.19).

3.4. Discussion

3.4.1. Comparing the performance of SSS-RAP using different resources

I compared the performance of SSS-RAP using data-sets generated with parameters from different resources.

When applying SSS-RAP to individual-level data of both quantitative traits (ECG) and binary outcomes (GBD), only part of the strong independent signals were detected compare to the conditional analysis results of meta-analysis. Flister et al suggested that single GWAS lack of power to test whether multiple causal SNPs underlie the same genomic region (Flister et al., 2013). The result of the current analysis support such suggestion.

As I have explained in the Chapter 3.2, the input data set of SSS-RAP contains four parameters: betas, standard errors of betas (SE), minor allele frequencies (MAF) and haplotype frequencies betas. Beta and SEs will normally be included in the meta-analysis summary results. However, MAFs and the haplotype frequencies (i.e. X_{11}) will rarely be provided by the meta-analysis, especially the haplotype frequency. So I mimicked three group level situation (Table 3-2), where the meta-analysis betas and standard errors were available, but MAF and haplotype frequency (X_{11}) need to be calculated from either internal (BWHHS) or external sources (HapMap). SSS-RAP results were closer to the conditional

analysis results when using BWHHS individual-level genotypes. On the other hand, this performance comparison also shown that results using HapMap individual-level genotypes as reference panel are less accurate than results using BWHHS genotypes as reference panel. The reason of such difference is because open access data (HapMap and 1000 Genome) have a limited sample sizes (less than 200 for each population) (<http://hapmap.ncbi.nlm.nih.gov/groups.html>) (<http://www.1000genomes.org/category/frequently-asked-questions/population>). Based on our simulation, we suggest the sample size of 1000 can lead us to an accurate LD estimation. Besides, in all cases, it is important that the allele and haplotype frequencies are from a population ancestrally matched to those in the meta-analysis.

3.4.2. SSS-RAP analysis of ECG traits

***NOS1AP* associated with QTc interval**

Numerous studies have reported association of common variation in *NOS1AP* with QTc interval (Newton-Cheh et al., 2009), and my fine mapping results suggest 3 independent effect SNPs. *NOS1AP* gene encodes a cytosolic protein that binds to neuronal nitric oxide synthase (nNOS). The protein has a carboxyl-terminal (C-terminal) PDZ binding domain. C-terminal PDZ binding domain interacts with NOS1 to accelerate cardiac repolarization by inhibition of L-type calcium channels in the heart and ventricular cells (Chang et al., 2008).

***SCN5A* associated with PR interval and QRS duration**

Using the meta-analysis results of the European population, I identified 4 and 5 possible independent effect SNPs in the voltage-gated, sodium channel type V, alpha subunit

(*SCN5A*) gene are associated with PR interval and QRS duration respectively, broadly consistent with previous reports (Pfeufer et al., 2009) (Newton-Cheh et al., 2009) (Pfeufer et al., 2010).

The two top SNPs with most significances, rs7372712 and rs7374540 for PR interval and QRS duration respectively, were not in LD ($r^2=0.02$ in HapMap Europeans). Rs7372712 is located in an LD block in the 5' region of *SCN5A*, which was replicated by the original meta-analysis (Gaunt et al., 2012). Additionally, rs7372712 is in very weak LD with the coding variants in *SCN5A* reported to influence ECG traits, specifically rs1805126 (D1819) and rs1805124 (H558R) in Europeans (HapMap $r^2=0.003$ and $r^2=0.011$, respectively). This finding replicated the claim that variations in the 5' end of *SCN5A* may influence on both PR interval and QRS duration.

The secondary independent SNP associated with PR interval is SNP rs12053903. It share a LD block with other previously reported SNP (rs6763048) effects in the 3' end of *SCN5A* (Bulter et al, 2012). The D' between these two SNPs are 1 and the r^2 between them are 0.068 (the 1000 Genome CEU population).

3.4.3. Meta-analysis of GIANT BMI dataset

Among all the significant SNPs I tested for BMI association using GIANT database, at least two independent effects have been observed previously for *MC4R* (Speliotes et al., 2010): top hit rs571312 and a coding SNP rs2229616 (V103I). In a separate gene-focused study, Loos *et al* (Loos et al., 2008) reported another *MC4R* SNP, rs17700633, as an independent effect SNP additional to the top hit, rs17782313 (HapMap $r^2=1$ to rs571312), for an

association of obesity. Rs17700633 is essentially in linkage equilibrium with the coding SNP rs2229616 in *MC4R*. These 3 SNPs were also detected by SSS-RAP program (Table 3-3), which again proves the accuracy of SSS-RAP program.

In addition, Speliotes EK *et al* (Speliotes et al., 2010) reported that several BMI-significant SNPs explained a substantial proportion of the association with the most significant SNP for the gene transcript in conditional analyses. These significant associations included *TMEM160*, *SLC39A8* and *SH2B1*. The SSS-RAP program did not identify any other independent SNPs except the top hit for these genes as well, which conformed the previous finding.

Fine mapping of ABCG5/8 region

GBD is a common disease in western countries, with more than 700,000 cholecystectomies performed in the U.S.A per year at a cost of ~ \$6.5 billion (Shaffer, 2006). This fine mapping analysis, using both LD pattern and SSS-RAP, shows that a multiple number of SNPs are independently associate with GBD risk in women of European ancestry. The pathway analysis shows that a huge amount of genotypes which associate with plasma lipid levels also associate with risk of GBD.

The leading SNP, rs4953023 ($P=3.87 \times 10^{-38}$, effect size=0.7003) is in perfect LD with the previously studied SNP rs11887534 (D19H in ABCG8). The coding variant D19H is believed to be a functional SNP which increases cholesterol secretion and hence the risk of super-saturation of bile (Katsika et al., 2010). The second SNP in *ABCG8*, rs4299376, tags an independent effect, probably reflecting a previously reported coding variant on GBD in Taiwanese, C54Y ($r^2=0.2$ between rs4299376 and rs4148211 in Europeans) (Liu et al., 2002).

Two other independent SNPs were identified in *ABCG5*, rs10439467 in intron 10 and rs6720173 (coding variant E604Q). The *ABCG5/ABCG8* transporter is the most proximal element of the cholesterol regulatory system to biliary cholesterol, and the identification of further new functional alleles at this locus emphasizes that any approach which might reduce cholesterol secretion without compromising TC and LDLc levels, could be useful in GBD prevention.

The comparison analysis of 63 independent loci for blood lipid levels and GBD, shows that a huge amount of SNPs influencing blood lipid levels influence GBD risk as well. I further analysis this finding separately by principal trait, LDLc, total cholesterol, HDLc or triglycerides. I found that a large number of loci influencing either LDLc or triglycerides, indicates the most prominent associations with GBD (rs4299376 in *ABCG5/ABCG8*; rs2081687 in *CYP7A1*; rs1260326 in *GCKR*; rs5756931 in *PLA2G6*; rs2479409 in *PCSK9* and rs1532085 in *LIPC*).

Relationship between correlation coefficient, beta and p value

I found it difficult to find a simple mathematical relationship between r^2 , beta and p value through the LD pattern of *ABCG5-8* region. But this piece of work became the start point of developing novel fine mapping methods using group-level data. The idea of utilizing LD pattern and processing stepwise selection were used to the development of SSS-RAP and another novel method I should explained in Chapter 5.

3.4.4. Conclusion

Firstly, when applied SSS-RAP to meta-analysis summary statistics, the results of SSS-RAP is highly consistent with the conditional analysis results, which proved its accuracy in

meta-analysis level. Moreover, I detected totally 12 SNPs (in *NOS1AP* and *SCN5A*) associated with QTc interval and PR interval. For the GIANT BMI meta-analysis, I found a suitable way to transform meta-analysis p values to meta-analysis betas and identified 17 regions with multiple association to BMI. For the *ABCG5-8* region associated with GBD, I identified four possible independent signals using both LD pattern and SSS-RAP. Last but not least, the idea of utilizing LD pattern and processing stepwise selection were used to the development of a novel fine mapping method which I should explain in the next chapter.

CHAPTER 4. HAPLOTYPE-BASED GROUP-LEVEL ASSOCIATION ANALYSIS USING AN E-M ALGORITHM

4.1. Introduction

In Chapter 3, I illustrated the limitation of GWASs and difficulties of fine mapping analysis in the current meta-analysis system. This motivated the development of the novel fine mapping method, SSS-RAP. It can detect SNPs with independent effects which are conditional on the strongest associated signal. The data used by SSS-RAP includes regression coefficients (betas), standard errors of the coefficient (SE) and minor allele frequencies (MAF) from meta-analysis and pairwise haplotype frequencies (X11) from a reference genotype panel (Zheng et al., 2013). I demonstrated its power and accuracy to detect secondary association signals in real life meta-analyses in Chapter 4. However, SSS-RAP is only applicable to pair-wise SNP comparisons and thus fine mapping multiple association loci at a group-level remains a problem.

4.1.1. Multiple regression

The classic method of testing multiple signals associated with a specific trait is multiple regression analysis (the term was first used by Pearson in 1908). The general purpose of multiple regression analysis is to understand the relationship between several predictor variables (also known as independent variables) and one criterion variable (dependent

variable). In the genetic fine mapping field, multiple regression is focused on testing the association between one trait and multiple genetic variants, especially SNPs. The regression model will be:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 \dots \beta_n * X_n + e \quad (0)$$

Where Y is the phenotype, $X_1 \dots X_n$ are the n covariates (SNPs) in the model. $\beta_1 \dots \beta_n$ are the beta coefficients of the n SNPs. “e” is the error of the regression model.

As explained in Chapter 1.3.2, single regression uses least squares approach to estimate the regression line. Multiple regression uses the same approach to find the regression line amongst multiple variants, but the regression line cannot be visualized in the two dimensional space.

Assumptions of multiple regression

There are several assumptions for multiple regression:

Firstly, the general assumption is that the relationship between SNPs is linear. This assumption can hardly be confirmed. Fortunately, multiple regression analyses are not mainly affected by minor deviations from this assumption.

Secondly, the residuals should be normally distributed. So it is necessary to check the distributions of the phenotypes of interest.

Thirdly, as illustrated in Equation (0), for a multiple SNPs model, it is quite common that any of the SNPs are strongly correlated with each other, or even if one SNP is in perfect LD with other SNPs. In this situation, the genotype matrix will suffer the wrong condition due to the presence of collinearity. The covariates will be confounded with one or more of the others and the precision of the estimation will become very poor.

The variance inflation factor (VIF) quantifies the severity of multi-collinearity in an linear least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity (Kutner et al, 2004).

Consider a linear model with k independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j + \dots + \beta_k X_k + \epsilon.$$

Where X_j is the covariance of j's variable of the linear model. VIF is defined as:

$$VIF = \frac{1}{(1 - R_j^2)}$$

Where R_j^2 is the multiple R^2 for the regression of X_j on the other covariates (a regression that does not involve the response variable Y). This identity separates the influences of several distinct factors on the variance of the coefficient estimate (Kutner et al, 2004). A large VIF value, for example bigger than 10, will indicate a collinearity problem in the regression model.

Finally, since the entire individual-level data is needed to process the multiple regression, it is not applicable for meta-analysis level data. Novel statistical methodology is needed to process approximate multiple regression using meta-analysis summary statistics directly.

4.1.2. Current state of art: GCTA conditional and joint effect analysis

To extend the scope of multiple regression to meta-analysis level, Yang J et al developed GCTA conditional and joint effect analysis (Yang et al., 2012). Since GCTA contains many functions, I only focused on its conditional and joint effect analysis, so GCTA in this chapter

only represent the GCTA conditional and joint effect analysis.

To process multiple regression at the group-level, GCTA separated the task into two steps. Firstly, GCTA built a linkage between the multiple regression and the single regression. It creates equations to convert the marginal effects (and SEs of marginal effects) to joint effects (and SEs of joint effects) using single SNP regression summary statistics and individual-level genotype data. Then for a meta-analysis of multiple studies, pooled individual-level genotypes of the whole discovery set are normally unavailable. So GCTA builds the approximate variance-covariance matrix of SNP genotypes using the allele frequencies from meta-analysis and LD correlations (r) between pair-wise SNPs from a reference panel. This variance-covariance matrix is then used by the least-squares approach to calculate the approximate joint SNP effects.

Assumptions and limitations of GCTA conditional and joint effect analysis

There are several assumptions for GCTA: firstly, the quantitative traits or disease outcomes are affected by multiple genetic variants. Secondly, the assumption of additive model of the genotypes was made. Moreover, the genotype matrix was adjusted by MAF. This is under the assumption that rarer variants have a larger effect size. In addition, the method also assumes HWE for all SNPs in the model. Finally, the method assumes that the LD correlation between a pair of SNPs should be similar in the meta-analysis and the reference panel (Ke et al., 2004).

For the final assumption, it has been demonstrated that combination of allele frequencies and pair-wise LD correlation may not always be biologically possible at the multiple-loci level,

since three-locus systems may place additional constraints on the maximum and minimum values for the pair-wise LD terms (Robinson et al., 1991).

4.1.3. Haplotypes and haplotype phasing

In Chapter 1, I gave a brief overview of haplotypes. In this chapter, I should describe how to utilize haplotype information and how to phase the haplotypes using statistical/computational tools.

In population genetics, haplotype information can be used in many analyses, for example, disease/quantitative traits mapping (Rieder et al., 1999) and inferring population histories (Harding et al., 1997).

Current genotyping methods such as microchip or next generation sequencing do not provide haplotype information. The cost of experimental haplotype phasing is very expensive (Browning and Browning, 2009). So haplotypes are normally phased by statistical algorithms. Two major algorithms have been used for haplotype phasing in this Chapter are Expectation-Maximization (E-M) algorithm and approximate coalescent and hidden Markov model (HMM) algorithm.

Since the novel method I developed is an E-M algorithm approach, I will explain the details of the E-M algorithm in the next section. For haplotype inferring, E-M algorithm is one of the earliest statistical approach (Excoffier and Slatkin, 1995). PLINK (Purcell S et al, 2007) contains a haplotype phasing function using an E-M algorithm. The E-M approach runs well for a small number of genetic variants, for example 10 SNPs. However, for a large numbers of markers, the E-M algorithm becomes time consuming and suffers a loss of accuracy by using suboptimal model for haplotype frequencies. Some examples of haplotype phasing software

using the E-M algorithm include Arlequin (Excoffier and Lischer, 2010) and PL-EM (Qin et al., 2002).

Moreover, the approximate coalescent model (McVean and Cardin, 2005) assumes that novel haplotypes are derived from old haplotypes by mutation and recombination, so that haplotypes in a small region tend to look similar. Furthermore, the HMM is a model in which the observations are conducted by an unobserved probabilistic process (Eddy, 2004). The coalescent-based methods and HMMs were applied by many statistical phasing approaches such as PHASE (Stephens et al., 2001), MACH (Li et al., 2010), IMPUTE2 (Howie et al., 2012) and BEAGLE (Browning and Browning, 2007). In each of these methods, the approximate coalescent are lead to an HMM, and the parameters are estimated using iterative algorithms such as E-M algorithm. Along with all of these HMM methods, MACH outperformed other methods for smaller sample sizes and BEAGLE outperform for large sample size (Browning and Browning, 2011).

As illustrated in Figure 4-1, for a specific population, the coinheritance of SNP alleles on the haplotypes leads to linkage disequilibrium amongst multiple loci (International HapMap Consortium, 2003). So theoretically, haplotypes are a better representation of LD than utilizing statistical correlation between pair-wise (r^2) SNPs.

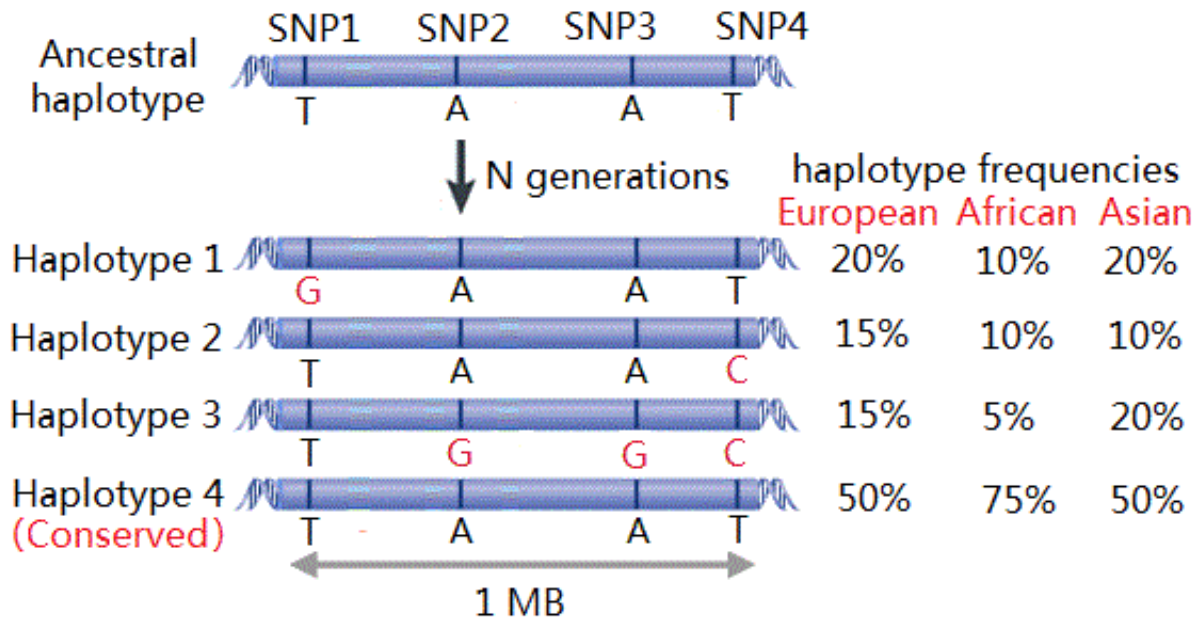


Figure 4-1. Example of haplotypes among four loci for different populations. In this example, four SNPs within 1 MB are selected. The above haplotype is the ancestral haplotype for these four SNPs. The four bottom haplotypes are the existing haplotypes of the ancestral haplotype after N generations. The haplotype frequencies are represented in the left side. As illustrated, the LD information for different populations is represented by the haplotypes.

4.1.4. Expectation-Maximization algorithm

In the statistical field, an E-M algorithm is an iterative method for detecting the value of the unobserved parameters that maximize the likelihood of the observed incomplete data (Dempster et al., 1977). Typically the models will involve latent variables except the unknown parameters and known observations.

The E-M algorithm can be applied in two respects. In the first case, if I cannot directly calculate the parameters directly due to the problems or limitations of the analysis, I have to use latent variables to link the observations and unknown parameters. In the second situation, I can calculate the unknown parameters but the latent variables can be simplified

with the likelihood function.

As illustrated in Figure 4-2, the E-M algorithm starts with setting random initial values for the unobserved variables. Then a numerical equation is built between the observed and unobserved variables. This process is called expectation (E step), which is usually the most difficult step of the E-M algorithm. The computational efficiency of the whole algorithm will highly depend on this step. Then I found the maximum likelihood of the expectation. This process is called maximization (M step). For the E step and M step, the expected values and the maximization do not need to be completed explicitly, but rather need only compute sufficient statistics for the completions. The E-M process will be kept until the latent variables converge to fixed points.

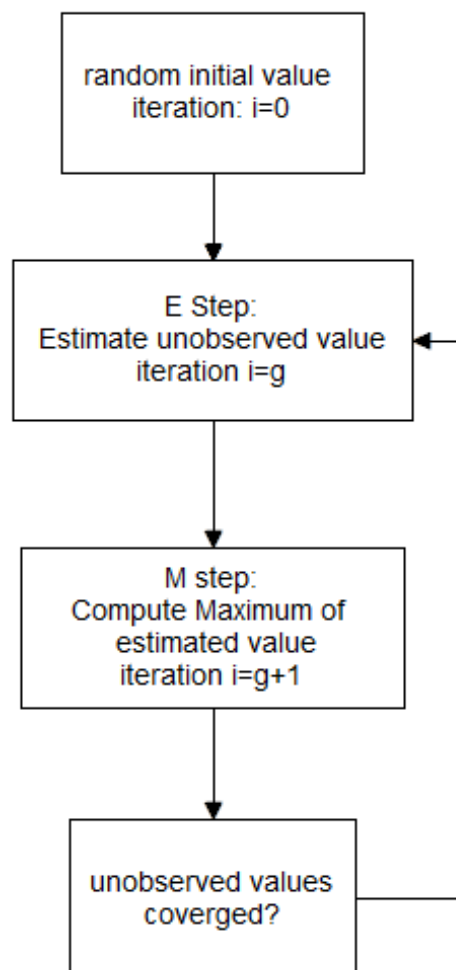


Figure 4-2. *An overview of the E-M algorithm.* After initialization, the E-step and the M-step are iterated until the unobserved values have converged. Iteration “i” means the iteration number.

Advantages and limitations of E-M algorithm

Compared to other optimization methods such as gradient descent (Yuan, 1999) or Newton-Raphson method (Ypma, 1995), the EM algorithm is simple and easy to implement. It does not usually require heavy preparatory analytical work and it is easy to program. Since the E-M algorithm is simple, it is computationally fast which is a very important characteristic for mining large datasets and only requires modest computer memory compared to other methods. In addition, the EM algorithm is numerically very stable. Moreover, the fitted values for the complete data are usually conducted during the E step, so there is no need of further calculation.

The E-M algorithm can only provide a local maximum of the objective parameter after converged. So it will always be helpful to process the E-M algorithm using multiple initial starting values to find the global maximum. For similar reasons, random initial parameters are useful to break symmetry in models as well.

4.1.5. Aims and Objectives

At the time I wrote this chapter, multiple regression (with a step-wise procedure) was still the gold standard of identifying SNP independence in individual-level. In meta-analysis level, the development of methods such as SSS-RAP (Zheng et al, 2013) and GCTA conditional and joint SNP effect analysis (Yang et al, 2012) fill the blank of process approximate conditional

analysis using meta-analysis summary statistics. The core concept of these methods are rebuilding LD structure using the allele frequencies in meta-analysis and LD correlations between pair-wise SNPs from a reference sample. However, I have a concern that utilizing allele frequencies and pair-wise LD correlation may lose LD information at the multiple-loci level since three-locus systems may place additional constraints on the maximum and minimum values for the pair-wise LD terms (Robinson et al., 1991). In contrast, haplotypes, which represent a specific sets of alleles observed on part of a chromosome, are the best way to represent LD among multiple loci. Fine mapping using haplotypes may pick up the LD information that is not detected using pairwise LD measures. However, it will be a challenge to develop such a fine mapping tool since the statistical system of haplotype is not as well established as the pair-wise correlation system. I should introduce an E-M style algorithm to utilise haplotype information as latent variables to link marginal SNP effects to joint SNP effects using meta-analysis summary statistics and develop the method using a E-M style algorithm (Dempster et al., 1977) definitely has the advantage of being simple, robust and easy to implement.

In this chapter, I shall describe the development of the E-M style algorithm haplotype regional association program (HAPRAP, pronounced (hap-rap)) software: a step-wise elimination program to detect the independent effect SNPs of a genomic region. The joint effects of the independent SNPs will be conducted after the best model is decided. The method shall be applicable to both individual-level data and meta-analysis summary statistics. When applying HAPRAP to meta-analysis summary statistics, the haplotypes will be phased from reference panel with individual-level genotypes. In addition, the method

shall be applicable to both quantitative traits and binary (disease) outcomes.

Objectives

- Develop a haplotype-based fine mapping method using an E-M style algorithm in complex quantitative / disease traits meta-analysis
- Compare the performance of the algorithm against the performance of existing fine mapping methods using sample datasets from 1) BWHHS individual-level data, 2) 1000 Genome individual-level data, 3) simulated meta-analysis data.

4.2. Methods and Materials

4.2.1. Overview of the methodology

The article presents an empirical method, HAPRAP, to estimate partial SNP effects using meta-analysis summary results. Given the assumption that haplotypes and haplotype frequencies in the reference panel are the same as those in the GWAS meta-analysis, HAPRAP uses haplotype effects to transform summary level marginal SNP effects to partial SNP effects.

To achieve the goal, HAPRAP has four steps (Figure 4-3): firstly, we initialize, generating random starting values for the partial SNP effects $J^{(0)}(S)$ and resulting haplotype effects $Q^{(0)}(H)$ (step 1). In step 2, we built an equation to adjust marginal SNP effects $E^{(g)}(S)$ based on haplotype effects $Q^{(g)}(H)$ and haplotype frequencies $p(H)$. In step 3, the estimated haplotype effects $Q^{(g)}(H)$ are themselves adjusted based on the deviation between the adjusted marginal SNP effects $E^{(g)}(S)$ and the observed marginal SNP effects $O(S)$. The

program iterates between Step 2 and Step 3 until the adjusted marginal SNP effects $E^{(g)}(S)$ are equal to the observed marginal SNP effects $O(S)$ within 10 decimal places. In the last step, we inverse transform the final estimated haplotype effects $Q^{(g)}(H)$ to the partial SNP effects $Jo^{(g)}(S)$. The notation of the HAPRAP algorithm is described in Table 4-1.

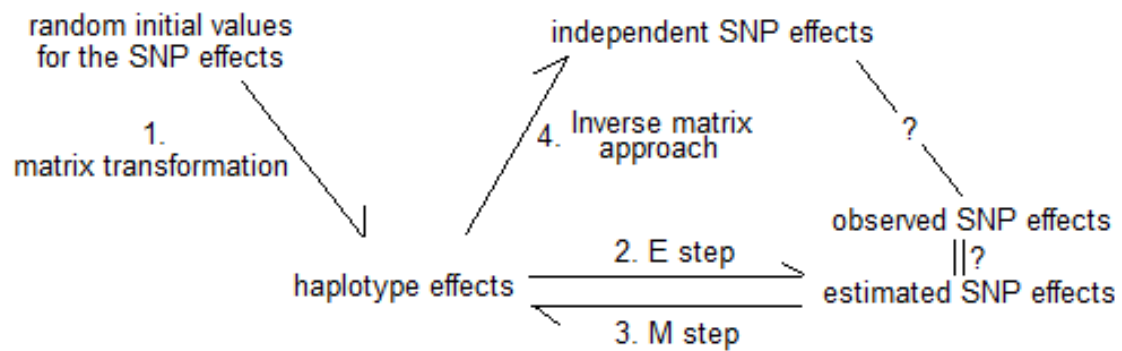


Figure 4-3. Schematic diagram of HAPRAP.

Assumptions

1. An additive model accounts for haplotype effects. For each SNP, the common allele is assigned as the baseline allele (with respect to trait); the rare allele is assigned as the “effect” allele.
2. The reference panel represents the same LD structure as the meta-analysis sample

Definitions

Constant vectors and matrix –observed data and inputs

$S = \{S_j\}$ is the set of j SNPs.

$H = \{H_n\}$ is the set of n haplotypes.

M represents the $n \times j$ haplotype matrix, where $M = \{M_{nj}\}$.

$p(H)$ represents the $n \times 1$ vector of the haplotype frequencies of the n haplotypes.

$O(S)$ represents the observed marginal SNP effects from GWAS / meta-analysis

Random initial values for the iteration

$Jo^{(0)}(S)$ is a $j \times 1$ vector which represents the random initial partial SNP effects of the j SNPs.

$Q^{(0)}(H)$ represents a $n \times 1$ vector of the random initial haplotype effects of the n haplotypes.

Variables in the g iteration – output for the g iteration

$E^{(g)}(S)$ represents the $j \times 1$ vector of the adjusted marginal SNP effects of the j SNPs in the g iteration. Whereas g represents the number of the current iteration and $g+1$ is the next iteration.

$Q^{(g)}(H)$ represents the $n \times 1$ vector of the adjusted haplotype effects of the N haplotypes in the g iteration.

$Jo^{(g)}(S)$ represents the partial SNP effects for the j SNPs on the g iteration.

$X(g)$ is the SNP with the greatest deviation between the observed marginal SNP effect and the estimated marginal SNP effect in the g iteration

Table 4-1. Notation of HAPRAP algorithm

Step 1. Initial values for partial SNP effects and transformation to haplotype effects: We set up a random set of initial partial effects for ‘j’ SNPs in the model, denoted $\mathbf{J}^{(0)}(\mathbf{S}_1)$, $\mathbf{J}^{(0)}(\mathbf{S}_2)$... $\mathbf{J}^{(0)}(\mathbf{S}_j)$. They are then transformed to the initial haplotype effects of the ‘n’ haplotypes, $\mathbf{Q}^{(0)}(\mathbf{H}_1)$, $\mathbf{Q}^{(0)}(\mathbf{H}_2)$... $\mathbf{Q}^{(0)}(\mathbf{H}_n)$, using the following matrix multiplication:

$$\mathbf{M} * \mathbf{J}^{(0)}(\mathbf{S}) = \mathbf{Q}^{(0)}(\mathbf{H}) \quad (1)$$

Where \mathbf{M} represents the $n \times j$ haplotype matrix. These initial haplotype effects are used as starting values in the first cycle of the iteration.

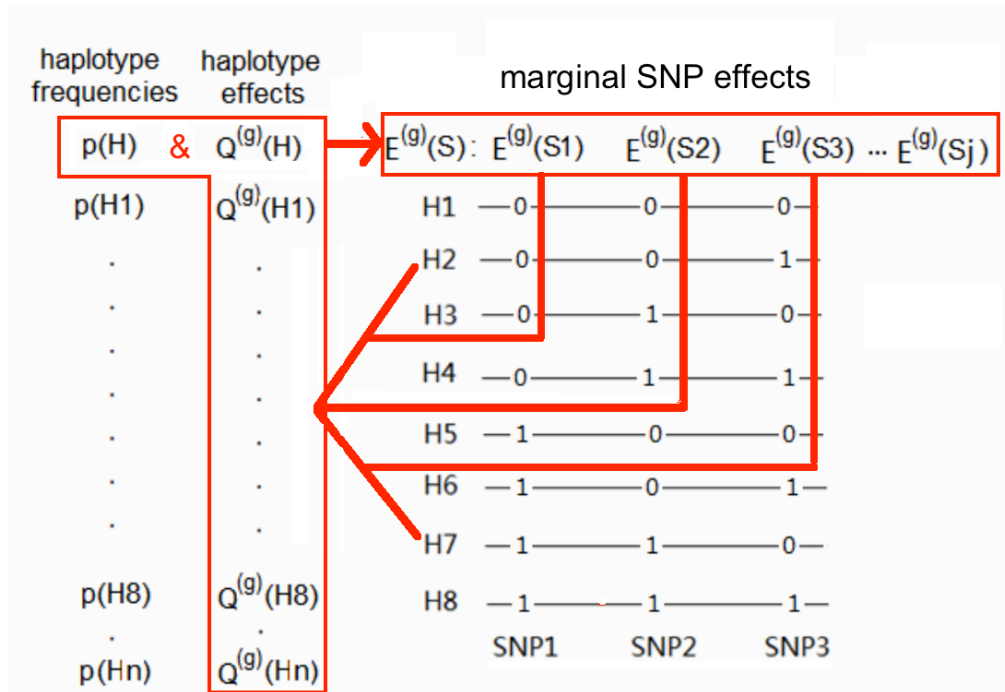


Figure 4-4. The SNPs-haplotypes matrix for HAPRAP. The iteration of HAPRAP is built based on a matrix summarising the haplotypes and haplotype frequencies for a certain population. “0” in the matrix means the haplotype contains the common allele for the relevant SNP, whereas “1” means the haplotype contains the rare allele for the relevant SNP. The small arrow (from left to right) is the marginal SNP effects estimation step shown in equation (2). The large arrow (from right to left) is the

haplotype effects adjustment step shown in equation (3) (4) and (5).

Step 2. The marginal SNP effects adjustment step: In a typical allelic association analysis (additive model), we treat the common allele as the “baseline” allele (with respect to trait) whereas the rare allele is the “effect” allele. The marginal SNP effect is the mean difference in phenotypes between the “effect” and “baseline”. Extending this idea to an additive haplotype model, for a certain SNP (called SNP A), we can divide haplotypes into two groups: “rare allele haplotypes”, H_{rare} , are haplotypes containing the rare allele of SNP A, which have real effects on the trait; “common allele haplotypes”, H_{common} , are haplotypes containing the common allele of SNP A, which represent the baseline. As in the single-SNP model, the difference in between the phenotypic mean in individuals with haplotypes containing the rare allele (the rare haplotype effect) and the phenotypic mean in those with the common allele (the common haplotype effect) is calculated. We get:

$$\text{Marginal SNP effect of SNP A} = \text{Effect of } H_{rare} - \text{Effect of } H_{common}$$

We adjust the haplotypic effects in each group to account for their relative frequencies. For H_{rare} , the adjustment is as follows: 1) Rare haplotypes with (relatively) higher frequencies are expected to have larger influence, each haplotype effect (Q) is weighted by its frequency. 2) The weighted rare allele haplotype effects are summed. 3) The weighted sum of the rare allele haplotype effects is then normalised by the sum of the frequencies of the rare allele haplotypes. The effect of H_{common} is constructed in the same way.

In HAPRAP iteration “g”, the marginal SNP effect, $E(g)(S)$, is adjusted by taking into account the haplotype effects, $Q(g)(H)$ and the haplotype frequencies $p(H)$. For a certain SNP ‘i’, the haplotypes are divided into two groups: $Q(g)(H_{rare})$ and $Q(g)(H_{common})$. The adjusted

marginal SNP effect of SNP “i”, $E(g)(S_i)$, is computed by taking the difference between average effects from the two haplotype groups:

$$E^{(g)}(S_i) = \frac{\sum Q^{(g)}(H_{rare}) * p(H_{rare})}{\sum p(H_{rare})} - \frac{\sum Q^{(g)}(H_{common}) * p(H_{common})}{\sum p(H_{common})} \quad (2)$$

We tested the reliability of equation (2) by a simulation and found that given any set of partial SNP effects, application of equation (2) never generated nonzero effect estimates for SNPs that were simulated to have truly null effects (Text S2).

Step 3. The haplotype effects adjustment step: In this step, the adjusted marginal SNP effects $E(g)(S)$ are compared to the observed marginal SNP effects (called $O(S)$). Reconciling the difference between $E(g)(S)$ and $O(S)$ is important because it equates the marginal SNP effects observed from the meta-analytic data with those that would arise under the distribution of haplotypes in the reference panel. The SNP-note as $X(g)$ -with the greatest deviation is adjusted for the next iteration $g+1$:

$$E^{(g+1)}(S_{X(g)}) = E^{(g)}(S_{X(g)}) - O(S_{X(g)}) \quad (3)$$

The other SNP effects remain the same to the next iteration:

$$E^{(g+1)}(S_i) = E^{(g)}(S_i) \text{ Where } i \neq X(g) \quad (4)$$

Then the haplotype effect $Q(g+1)(H)$ for SNP $S_{X(g)}$ will be weighted based on the above change and be used in the next iteration:

$$Q^{(g+1)}(H_{rare}) = Q^{(g)}(H_{rare}) + E^{(g+1)}(S_x) \quad (5)$$

Step 4. Convergence and the inverse matrix approach: After the adjusted marginal SNP effects, $E(g)(S)$ are approaching the observed SNP effects, $O(S)$, within 10 decimal places, $Q(g)(H)$ are inverse transformed to j partial SNP effects using the generalized inverse matrix

multiplication:

$$Y^{-1} * Q^{(g)}(H) = J_o^{(g)}(S) \quad (6)$$

We call these estimated partial SNP effects $J_o(g)(S)$ “HAPRAP betas”.

4.2.2. Estimating the standard error of HAPRAP betas using simulation based HAPRAP

Pre-test of SNP significances

Generating bootstrap standard errors can use computational resources intensively. To improve computational efficiency, we first pre-test the significance of the candidate SNPs using the HAPRAP betas and the standard errors of the meta-analysis betas (since the uncertainty of the effect of a given SNP is larger in a multivariate model than that in a single SNP model). SNPs with the highest p value will be step-wise eliminated from the model until all SNPs reach the p value threshold we set.

For a model with multiple SNPs, the standard error of a specific SNP taken from an individual regression analysis is usually smaller than the multiple regression standard error for the same SNP, since multiple SNP models will create more uncertainties than signal SNP ones. Based on this assumption, a pre-significant test was processed to improve the computational time of HAPRAP.

The pre-significant test uses HAPRAP betas and the standard errors of the observed SNP effects to conduce the t-statistics:

$$T \text{ statistic} = \frac{\text{HAPRAP beta}}{\text{standard error of observed SNP effects}}$$

The t-statistics were transformed to t test p values using R code:

$$2 \times \text{pt}(-\text{abs}(t\text{-statistics}), \text{df}=N-1)$$

Where pt is the code to convert t-statistics to p-value, “df” is the degree of freedom, N is the sample size.

After the t test p values were calculated, a hypothesis test is conducted to see whether the p value is smaller than the p value threshold, i.e. 5×10^{-8} , if so, it will be very rare that the SNP is significant by chance. So the null hypothesis is that the SNP is significant and will be accepted.

For the pre-significant test, the SNP with highest p value will be backwards eliminated from the model until every SNP in the model reaches the p value threshold after multiple testing correction. The problem of such pre-significant test is that the p value here will be more stringent than it should be so some SNPs with p values near the threshold will be eliminated.

To calculate the standard errors of the remaining SNPs in the model after the pre-significant test, I developed a simulation based HAPRAP program (simHAPRAP).

Population simulation using simHAPRAP

As illustrated in Figure 4-5, the simulation starts by simulating a population with sample size equal to the total number of participants in the studies / meta-analysis. Genotypes are randomly generated based on the haplotypes and haplotype frequencies. For a quantitative trait, phenotypes are selected from a normal distribution with mean equal to zero and standard error equal to the deviation of the trait. For a disease outcome, binary phenotypes are selected from a binomial distribution which fits the probability of cases and controls of the trait.

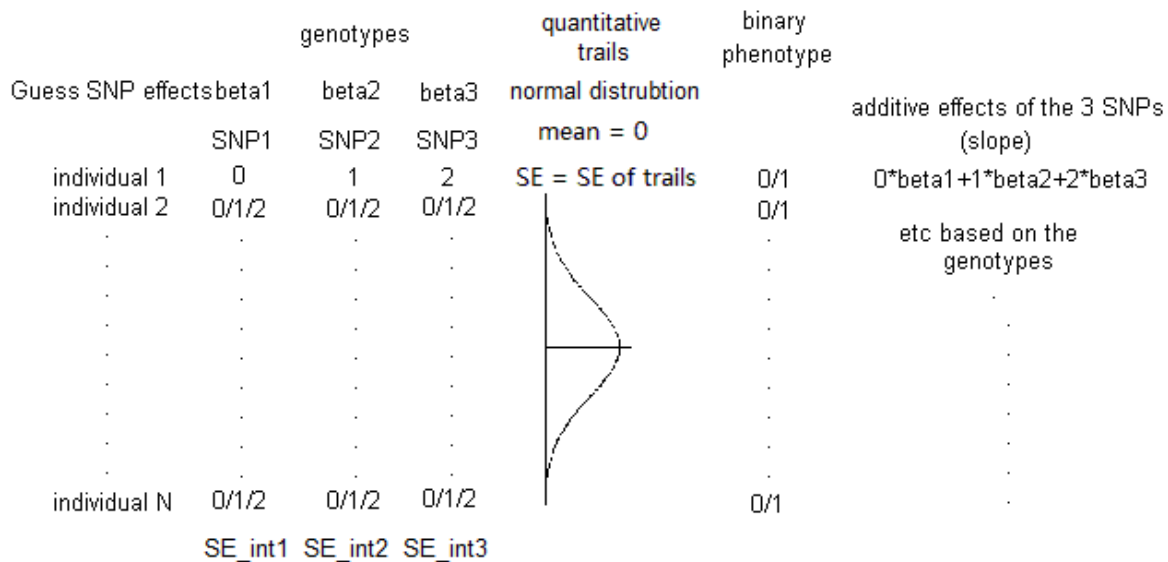


Figure 4-5. Population simulation for simHAPRAP.

SimHAPRAP program

As illustrated in Figure 4-6, the simHAPRAP program is similar to HAPRAP except for several differences: Firstly, since genotypes and phenotypes are simulated, the haplotype effects are no longer used as latent variables. The random initial values of the marginal SNP effects are used directly to estimate joint SNP effects. The E-M algorithm is processed between estimated joint SNP effects and estimated marginal SNP effects. This means the E-step become linear regressions between the phenotype and genotype of each SNP respectively, whereas the M-step remains the same. After the estimated joint SNP effects converge, these effects are the optimal joint / independent SNP effects.

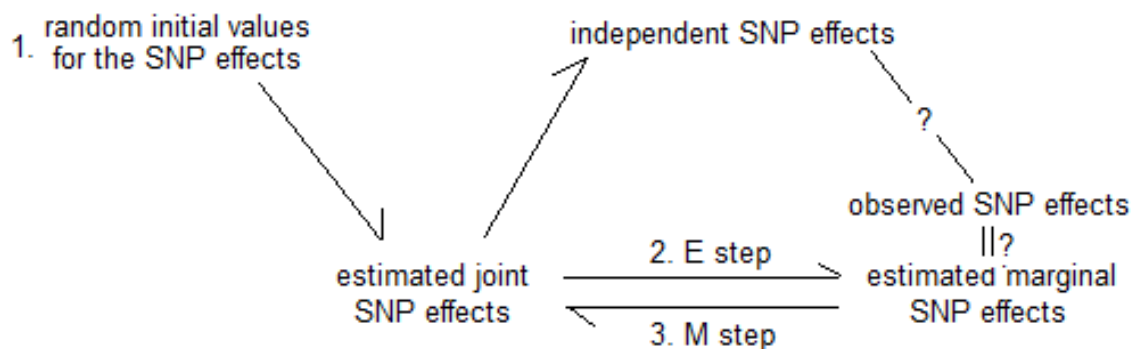


Figure 4-6. *Flowchart for simHAPRAP.*

Stepwise backwards elimination

The t-statistics are calculated using HAPRAP betas and simHR SEs, for each remaining SNP in the model:

$$T \text{ statistic} = \frac{HAPRAP \text{ beta}}{\text{simHR standard error}}$$

P values are transformed using the same way explained above. The SNP with the highest p value will be eliminated one by one until the optimal model is selected.

4.2.3. Flowchart of HAPRAP program

Figure 4-7 illustrates the workflow of the HAPRAP program in. The core of the program is the E-M algorithm. Other methods used in the program include: haplotype phasing tools, significant test, step-wise procedure and simHAPRAP program.

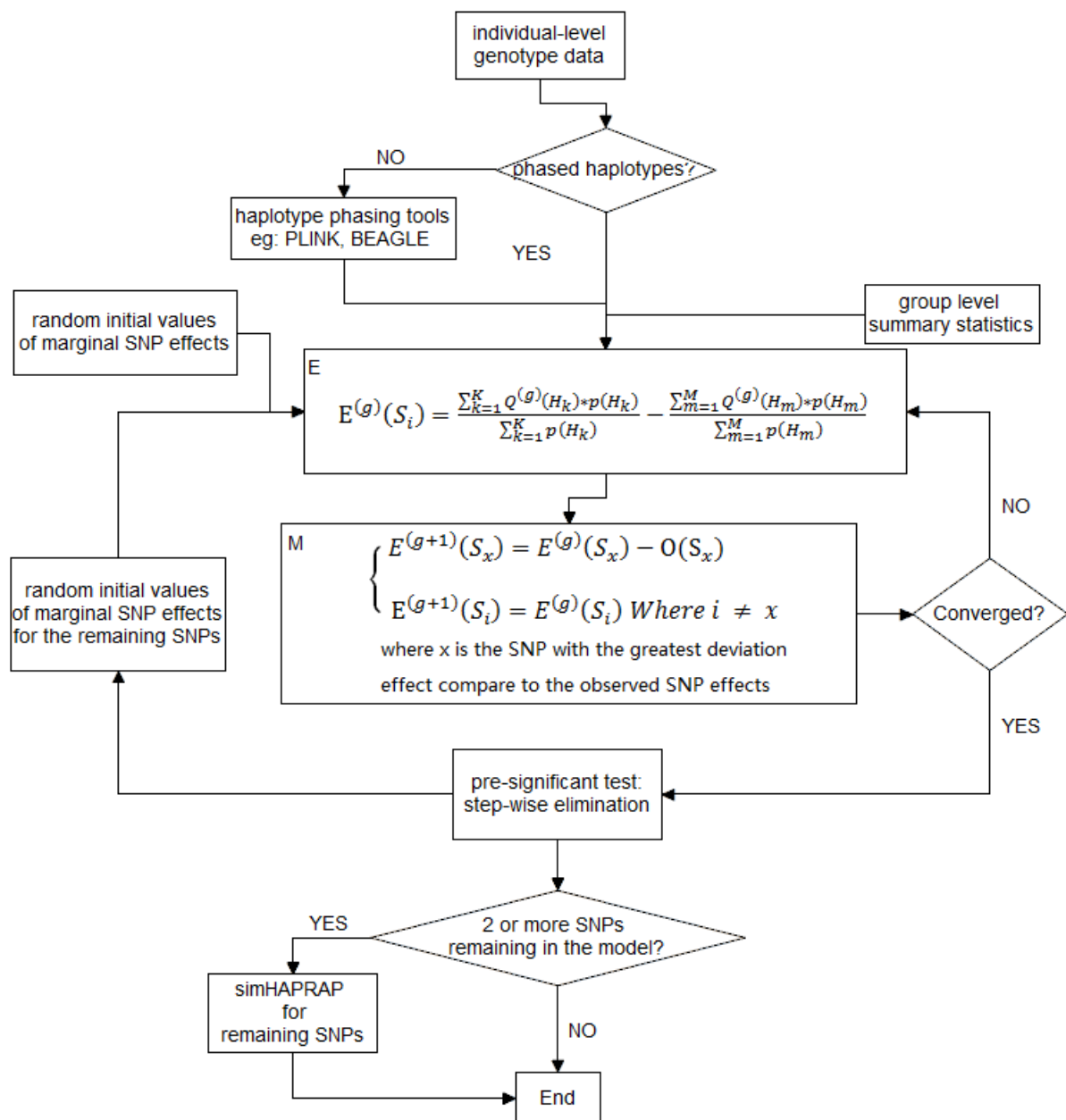


Figure 4-7. Workflow of HAPRAP program.

4.2.4. HAPRAP web-based tool

A web-based implementation of the HAPRAP program was developed using HTML and CSS language (Figure 4-8). The source code for this program is available to academic users on application to the authors. The webpage has been tested on an Ubuntu server running Apache 2 Web-server. The URL of HAPRAP home page is

<http://apps.biocompute.org.uk/haprap>

HAPRAP Home download

E-M algorithm Haplotype regional association program (V1.0)

The E-M algorithm HAPlotype Regional Association analysis Program (HAPRAP) is a fine mapping method which uses group level summary statistics from a meta-analysis of GWASs and haplotypes from an individual-level reference panel.

Download

To use HAPRAP, please [download](#) the HAPRAP package. Then unzip the file to your target directory, for example:

```
tar -zxvf HAPRAP.tar.gz
```

Important:

1. please put all five files together in one folder.
2. please install [PLINK](#) before using HAPRAP
3. The HAPRAP program is only working in python 2.7 or above

Figure 4-8. HAPRAP web-based tool

4.2.5. Sample datasets

BWHHS individual-level dataset

As described in the Chapter 2, 1980 BWHHS samples, which were genotyped by Illumina Cardio-MetaboChip were used for performance comparison (Shah et al., 2013b). I selected three genes associated with ECG traits and 14 genes associated with blood plasma traits. All the regions and associated traits are listed in Table 4-2. In total, 117 SNPs were selected within these regions. To prepare the input dataset for GCTA and HAPRAP, signal SNP model regression was conducted in R (code `lm`). MAF of the 117 SNPs were estimated using PLINK.

Genomic regions	Associated traits
<i>LIPC</i> and <i>LPL</i>	HDLc
<i>APOE</i> , <i>PVRL2</i> , <i>TOMM40</i> , <i>LDLR</i> , <i>SMARCA4</i> , <i>CELSR2</i> and <i>APOB</i>	LDLc

<i>APOE, LDLR, SMARCA4, PVRL2 and APOB</i>	TC
<i>NOS1AP</i>	QTc interval
<i>SCN5A and SCN10A</i>	PR interval

Table 4-2. *List of phenotypes and genotypes selected from BWHHS.*

To avoid the collinearity of the above multiple SNPs models, VIFs within each of the above genomic regions were estimated using R package 'car' (<http://www.r-project.org/>). All SNPs with VIF higher than 10 were eliminated one by one from the regression model until all remaining SNPs reached the VIF threshold.

I phased the BWHHS unphased individual-level genotypes using PLINK. The haplotype letter (A, G, C or T) format was converted to numeric format (0 or 1) format using a python script I conducted (the script was attached in Chapter 9.5).

1000 Genome individual-level dataset

I used 1000 Genome individual-level data here for performance comparison. To increase the sample size, I combined 1000 Genome CEU unrelated individuals and British in England and Scotland (GBR) individuals. From the 117 selected SNPs, I removed SNPs not existing in 1000 Genome and SNPs with VIF higher than 10. After filtering, 67 SNPs remained for the performance comparison. The principal-component analysis (PCA) of the 67 SNPs is illustrated in Figure 4-9. The PCA suggested that there is no population stratification between the CEU, GBR and BWHHS populations

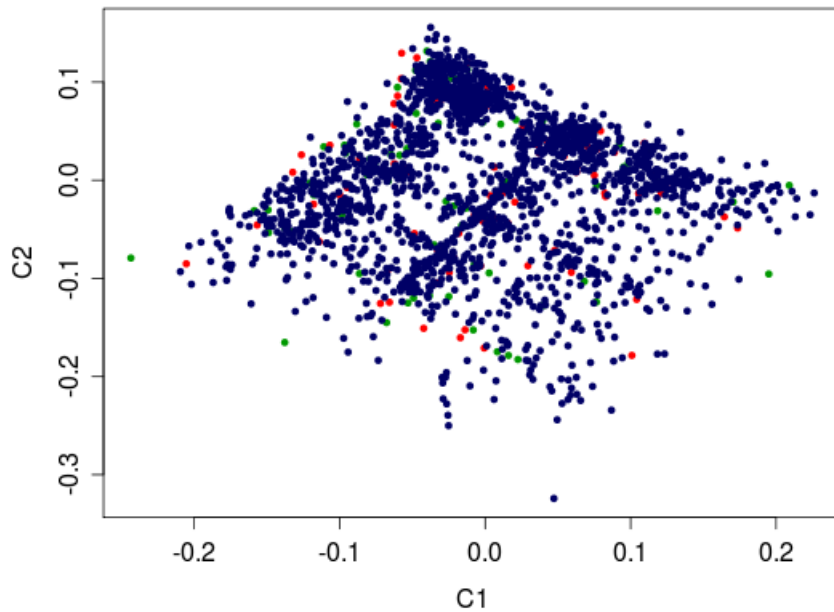


Figure 4-9. *The Multidimensional Scaling Plot of 1000 Genome CEU and GBR Population*

Together with BWHHS Individuals. Points in green are from the CEU population, points in red are from the GBR population and points in blue are from BWHHS. The three populations cluster together which suggests no population stratification between the populations.

Simulations of artificial meta-analysis

To conduct a performance comparison among existing methods and HAPRAP, simulations are needed.

I conducted 2-SNPs models and 3-SNPs models for this simulation. There is no need to build any model containing more than 3 SNPs since no additional constraints on pair-wise disequilibrium values exist in a four loci situation, except those executed by the three loci combinations containing the two loci of interest (Robinson et al., 1991).

The simulation starts with setting the parameters. To mimic complex three SNP models here, I firstly assumed that SNP1 is a top hit with independent effect of 1. SNP2 was set as

negative control without any independent effect. I used it to test whether the method can eliminate the bystander SNP. Moreover, SNP3 was set as a secondary signal with a small independent effect of 0.3. I set this effect size to test whether HAPRAP can detect this small effect even if the secondary SNP (SNP3) is in high LD to the top hit (SNP1). Since this approach is designed for common variants analysis, I set the MAFs as 0.4, 0.38 and 0.3 respectively for SNP 1, 2 and 3 so that it will be easier to set different LD correlations between these SNPs.

Using the same way described in Chapter 5.2.2 simHAPRAP section, I simulated the genotypes and phenotypes for 100,000 individuals for each model (Table 4-3) as a meta-analysis of 100,000 people. To calculate the marginal SNP effects (with standard errors), I conducted signal SNP model association test for each of the SNP. The MAF counts were estimated using PLINK. To mimic the situation that individual-level data sharing is unavailable for this meta-analysis. I only used meta-analysis summary statistics: the marginal SNP effects (with SE) and MAFs here. To test whether sample size of the reference genotype panel will affect the results, I randomly picked 10000, 5000, 1000, 500, 200, 175 or 150 individual-level genotypes from the population pool. I used them as reference panel. The haplotypes were phased using PLINK function:

```
plink --file data --hap mySNP.hlist --hap --freq
```

“data” file is the plink format data.ped and data.map file maintaining genotype information and mySNP.hlist as the following format:

```
* SNP1 SNP2 SNP3
```

Where the first character is an asterisk*, followed by a tab-delimited list of SNPs in the same

order as data.map.

A. 2 SNPs Model		
model	r ² between SNP1 and SNP2	
1	0.9	
2	0.8	
3	0.5	
4	0.2	

B. 3 SNPs Model		
model	r ² between SNP1 and SNP2	r ² between SNP 1 and SNP3
1	0.8	0.5
2	0.8	0.3
3	0.8	0.1
4	0.5	0.5
5	0.5	0.3
6	0.5	0.1

Table 4-3. Assumptions of Pair-Wise LD (r^2) of SNPs Models for Population Simulations. **A**, four 2-SNPs models with different pair-wise r^2 between SNP1 and SNP2. **B**, six 3-SNPs models with different pair-wise r^2 between SNP1 & SNP2 and SNP1 & SNP3.

4.2.6. Performance comparisons

I tested the performance of HAPRAP against other existing methods, i.e. multiple regression (Generalized Linear Models coded as “glm” in R) and GCTA joint SNP effect analysis (<http://www.complextaitgenomics.com/software/gcta/index.html>). These approaches were applied to the following datasets.

BWHHS individual-level data

I applied multiple regression, GCTA version 1.2 and HAPRAP to BWHHS individual-level data.

BWHHS summary statistics (marginal SNP effects, standard errors and MAFs) of the 117 SNPs and individual-level genotypes are used to process GCTA and HAPRAP. For multiple regression, which is the most established method for fine mapping, the joint SNP effects were estimated using the whole individual-level phenotypes and genotypes. For this reason, the results of multiple regression were assumed to be the “gold standard” for this comparison. I conducted two two-tailed t tests: the first one compared the HAPRAP betas to gold standard and second compared GCTA betas to gold standard. I drew histograms to check the distribution of these two sets of t-statistics. To compare the accuracy of HAPRAP and GCTA, I drew a QQ plot for the sorted and unsorted t-statistics.

Moreover, I calculated the concordance correlation coefficients (noted as CCC) (Lin and Torbeck, 1998) between HAPRAP betas (or GCTA betas) and multiple regression joint SNP effects using a function in R package called “epiR” (<http://cran.r-project.org/web/packages/epiR/epiR.pdf>). CCC measures the agreement between two methods. It combines measures of both precision and accuracy to determine how far the betas deviate from the line of perfect concordance (the $y=x$ line) 23. The method (HAPRAP or GCTA) with CCC estimates were closer to 1 got More tightly correlated with the gold standard. In addition, pair plots were drawn between each pair of betas to compare the agreement between HAPRAP (GCTA) and multiple regression using a famous R graphic package “ggplot2” (<http://cran.r-project.org/web/packages/ggplot2/index.html>).

BWHHS summary statistics and 1000 Genome individual-level data

I further compared the performance of HAPRAP and GCTA in the circumstance that only external genotypes are available. I used the BWHHS summary statistics (marginal SNP effects, standard errors and MAFs) of 67 SNPs and selected 1000 Genome CEU and GBP individual-level genotypes as reference panel. The same t test described in the above paragraph was conducted. Histograms and QQ plots were drawn for two sets of t-statistics as well. Scatter plots, CCC estimates and pair plots comparing multiple regression joint effects and HAPRAP betas or GCTA betas were conducted.

An independent benchmark using simulated meta-analysis data

I then compared the performance of HAPRAP and GCTA using the artificial meta-analysis I described in Chapter 5.1.5.

In fact, the independent SNP effect I set (1, 0 and 0.3 for SNP1, 2 and 3 respectively) did not take into account the impact of sampling errors of the simulation, so I used multiple regressions joint SNP effects for the 100,000 individuals as gold standards in this comparison. For each model, I applied HAPRAP and GCTA using the summary statistics and the genotype reference panel for a specific sample size. Since multiple regression is not applicable to group-level data, so I applied multiple regression using individual-level phenotypes and genotypes from the reference panel. For each method, I repeated the process 1000 times. The mean and standard deviation of the 1000 replications were calculated. The difference between the mean of 1000 replications and the gold standards were noted as error of each method. The standard deviation of the 1000 replications was used to conduct the 95% CI.

Performance comparison for rarer variants using simulated meta-analysis data

To test the performance of HAPRAP in rarer variant analysis, I performed a simulation for two SNPs with $MAF < 0.1$. I assumed that SNP1 has an independent SNP effect of 1 and SNP2 is a bystander SNP without own effect. The minor allele frequencies of SNP1 and SNP2 are 0.093 and 0.083 respectively. The pair-wise LD (r^2) between the two SNPs is 0.9, which is extremely high. Sample size of the reference panel will be 10000, 5000, 1000, 500 and 100 respectively.

4.3. Results

4.3.1. The independent benchmark of HAPRAP, GCTA and multiple regression using simulated meta-analysis data

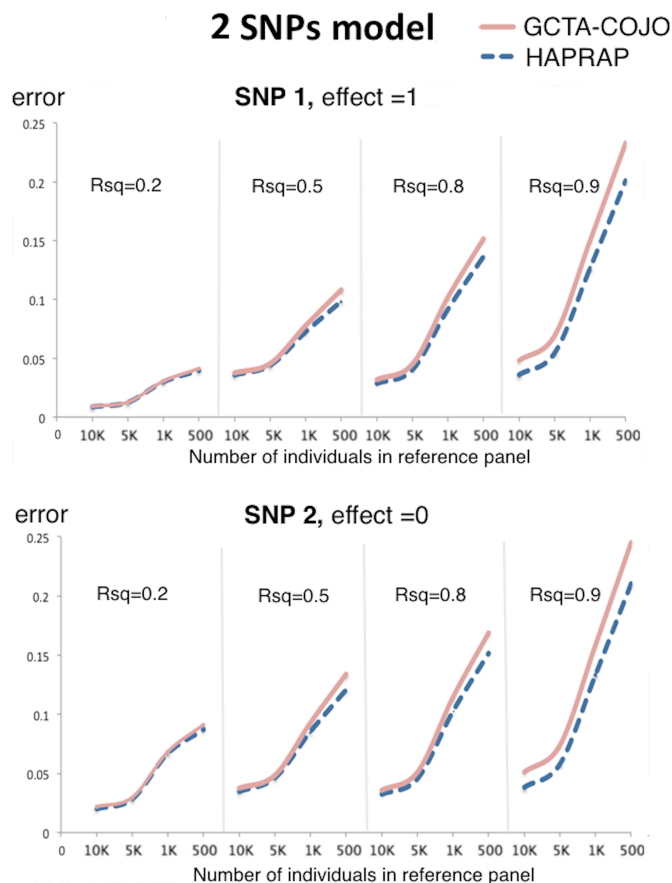
I compared the performance of three methods using simulated meta-analysis data. The comparison results were represented in Figure 4-10, Figure 8-1 and Table 8-8. These comparisons show that the sample size of the reference panel will affect the accuracy of all the methods. I found that with the increase of sample size of the reference genotype panel (N), the mean of the errors are nearer zero and the 95% CI of errors are narrower. A sample size of 500 or more will return relatively accurate results. When sample size is lower than 175, which is the approximate number of Caucasians in the 1000 Genome project, the errors deviated too widely and it will be risky to rely on the conclusion of any analysis. In addition, the correlation between two SNPs increases with the increase of their pair-wise r^2 . This

increased difficulty of eliminating the bystander SNP will cause the mean and 95%CI of errors to become more deviated and wider.

For the two SNPs models (Figure 4-10A, Figure 8-1A), the accuracies were similar between HAPRAP and GCTA when r^2 is low to modest. When r^2 is high (i.e. $r^2=0.8$ and $r^2=0.9$), HAPRAP outperforms GCTA in among different sample sizes of reference panels.

For the three SNPs models (Figure 4-10B, Figure 8-1B), HAPRAP outperforms GCTA in almost all kinds of simulations I did. HAPRAP has major advantages against GCTA to estimate the independent effect of the secondary signal (SNP3) when r^2 is high or modest.

When comparing HAPRAP and GCTA results to multiple regression results using individual-level data of the reference panel (Figure 4-10C, Figure 8-1C), HAPRAP and GCTA both have a more accurate mean and narrower error bar compared to multiple regression.



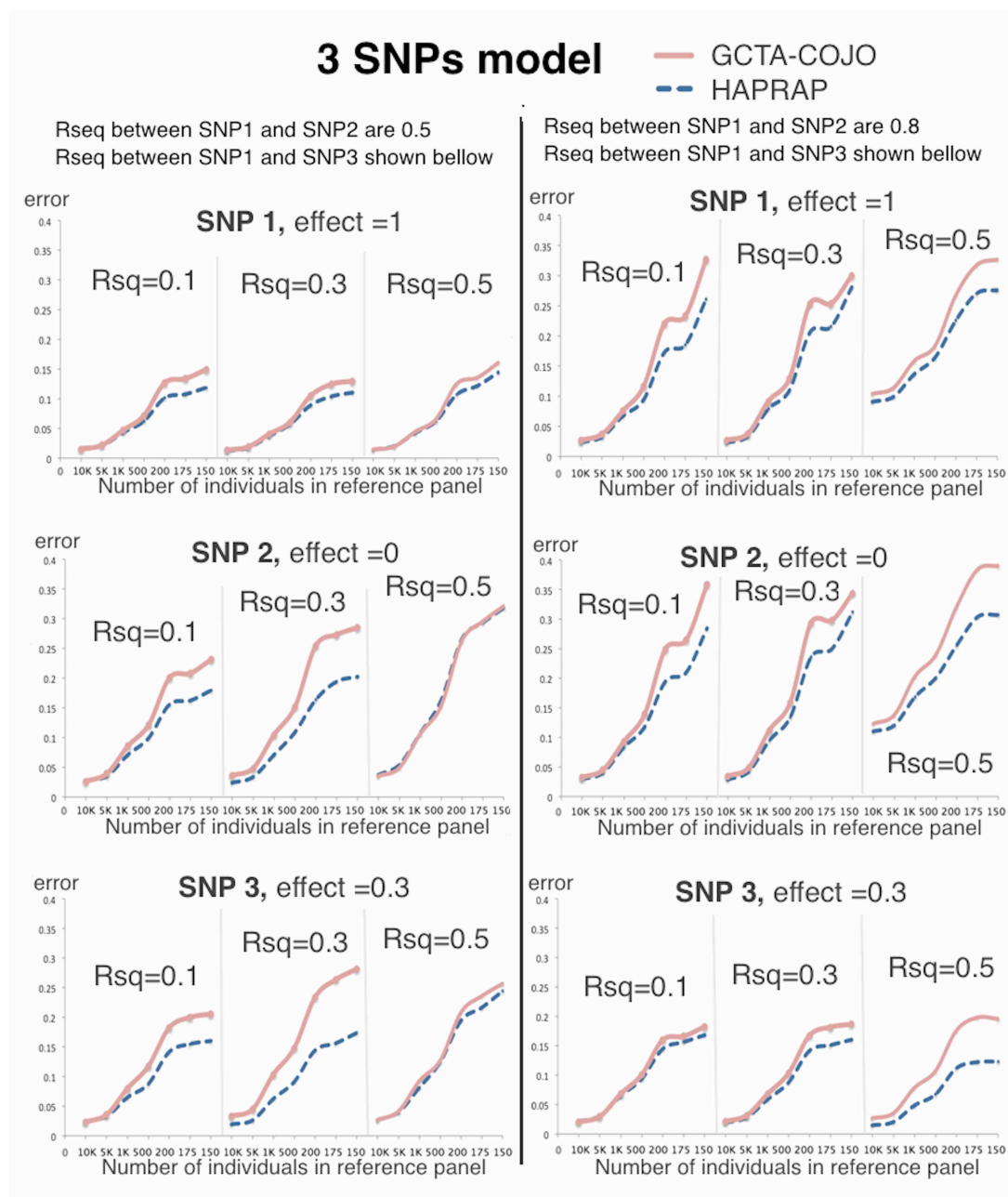


Figure 4-10. Performance Comparison of HAPRAP, GCTA and Multiple Regression Using Meta-Analyses of the Simulated Populations. Error (represented as a bar) is defined as the difference between the multiple regression joint SNP effect and the mean (with 95% CI) of 1000 replications. N is the number of individuals in the reference panel. (A) An example of the comparison between HAPRAP and GCTA for 2 SNPs with $r^2=0.8$. (B) The comparison between HAPRAP and GCTA for 3 SNPs, $r^2=0.8$ between SNP1 and SNP2, and $r^2=0.5$ between SNP1 and SNP3. (C) The comparison between HAPRAP, GCTA using meta-analysis data and

multiple regression using individual-level data from the reference panel. Two SNPs with r^2 of 0.5 were used here as an example.

Performance comparison of rarer variants

On the other hand, I compared the performance of HAPRAP against GCTA in rarer variants analysis. In the example of two SNPs with $MAF < 0.1$ and $r^2 = 0.9$, HAPRAP outperform GCTA in different sample sizes (Table 4-4).

SNP1						
$r^2=0.9$						
	HR_MEAN	HR_SD	Mreg_MEAN	Mreg_SD	GCTA_MEAN	GCTA_SD
10K	0.002579	0.011049	0.000874	0.239421	0.003993	0.03148
5K	0.003626	0.015617	0.023512	0.344624	0.001064	0.04715
1k	0.005342	0.038019	0.002425	0.804111	0.010605	0.112993
500	0.014124	0.055718	0.001902	1.15045	0.052179	0.217501
100	0.014404	0.122847	0.069349	2.682938	0.057968	0.292097
SNP2						
$r^2=0.9$						
	HR_MEAN	HR_SD	Mreg_MEAN	Mreg_SD	GCTA_MEAN	GCTA_SD
10K	0.002847	0.012574	0.001868	0.249512	0.004604	0.03517
5K	0.003952	0.017793	0.020453	0.359565	0.000881	0.052533
1k	0.005489	0.043123	0.012581	0.843051	0.01035	0.124852
500	0.014384	0.062825	0.013464	1.204385	0.054212	0.237283
100	0.013037	0.154193	0.050376	2.783836	0.054051	0.324903

Table 4-4. Performance Comparison of HAPRAP, GCTA and Multiple Regression Using a Simulated Population with Two Rare Variants. R^2 is the pair-wise LD r^2 . N is the number of individuals in the genotypes reference panel. HR_MEAN (and HR_SD) are the means (and standard deviations (SD)) for differences between HAPRAP betas and multiple regression joint SNP effects. Mreg_MEAN (and Mreg_SD) are the means (and SDs) for differences between multiple regression

joint SNP effects of the reference panel and multiple regression joint SNP effects of the whole population. GCTA_MEAN (and GCTA_SD) are the means (and SDs) for differences between GCTA betas and multiple regression joint SNP effects.

4.3.2. A performance comparison of HAPRAP against GCTA using BWHHS genotype data

I conducted a performance comparison between HAPRAP and GCTA using the BWHHS individual-level data for 1980 individuals. As illustrated in Figure 4-11 A and B, the t-statistics for both HAPRAP against multiple regression and GCTA against multiple regression are normally distributed. The means of both t-statistics were near zero. The most deviated t-statistics were around 0.3 for both methods. After I converted t-statistics to t test p values, the most deviated p values were 0.746 and 0.785 for HAPRAP and GCTA respectively (Table 8-9). This suggests that the results of HAPRAP and GCTA have no statistical differences compared to multiple regression results when individual-level data is available. As shown on the scatter plots in Figure 4-11 C and D, most of the t-statistics and p values were near the X=Y line, which means HAPRAP and GCTA are performing at approximately the same level when using the entire individual-level genotypes as a reference panel.

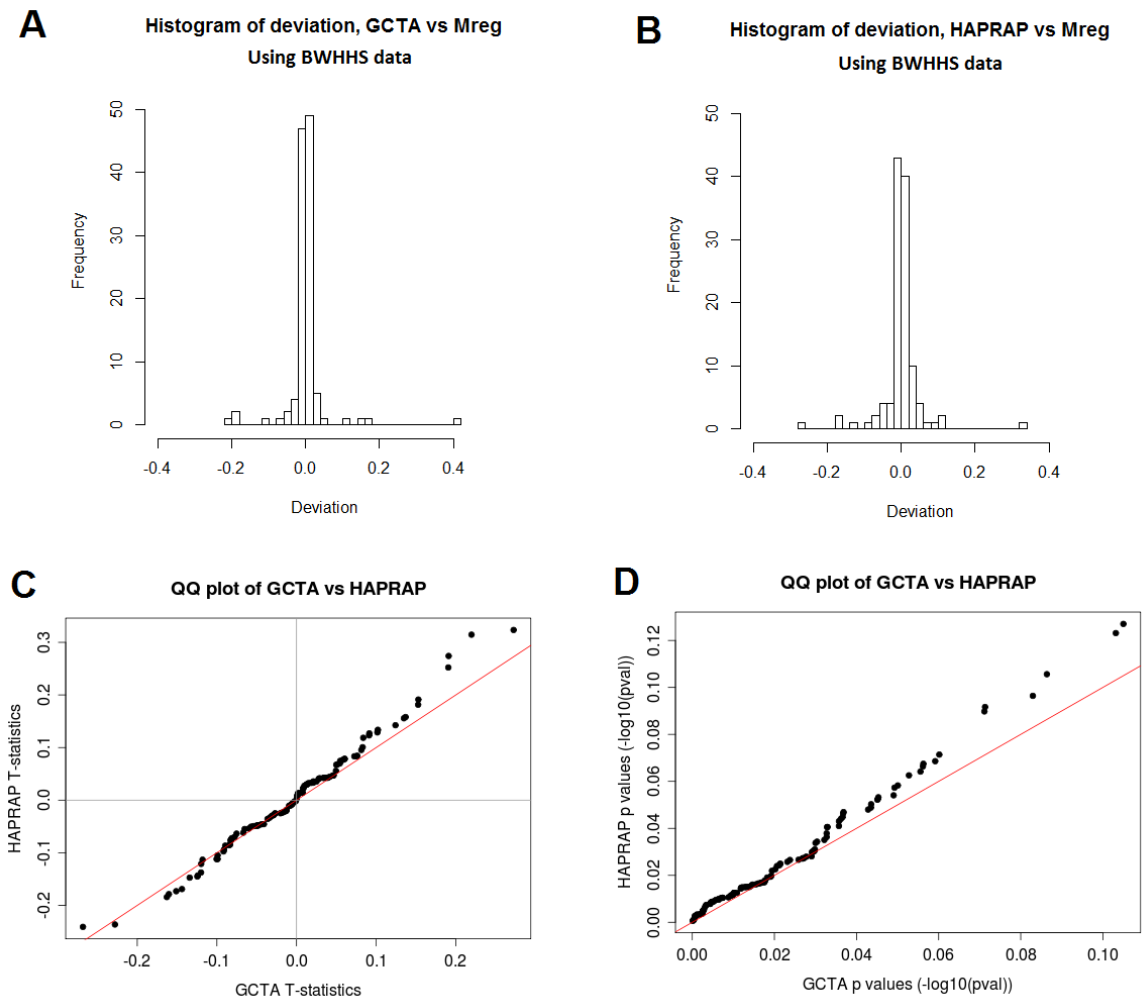


Figure 4-11. Performance comparison against HAPRAP and GCTA joint effect analysis using BWHHS unphased genotypes. I conducted a two tailed t-test to compare the difference between multiple regression (Mreg) joint SNP effects and HAPRAP / GCTA betas for all 117 SNPs inside the region listed in Table 8-9. The entire BWHHS individual-level genotypes and phenotypes were used for the comparison. **A** histogram of T-statistics for Mreg betas vs GCTA betas (called GCTA T-statistics). **B**, histogram of T-statistics for Mreg betas vs HAPRAP betas (called HAPRAP T-statistics). **C**, QQ plot for GCTA T-statistics against HAPRAP T-statistics. **D**, QQ plot of GCTA against HAPRAP using $-\log_{10}(P)$ as a scale. The red line in **C** and **D** is the X=Y line. Detailed results were listed in Table 8-9.

Moreover, as shown in Figure 4-12 A and B, all the points of the scatter plots were close to the $X=Y$ line, which suggests good correlations amongst the three methods. Mining further, I compared the concordance correlation coefficient analysis results. As illustrated in Table 4-5, HAPRAP had a slightly better agreement to multiple regression than GCTA (CCC estimate 0.9982 for HAPRAP and 0.9978 for GCTA). The bias correction factor of HAPRAP is also closer to 1 than that of GCTA (bias correction factor 0.9997 for HAPRAP and 0.9996 for GCTA). For the pair plots shown in Figure 4-12 C (and results in Table 8-9), I found that most of HAPRAP betas are very close to gold standards (multiple regression joint SNP effects), whereas some of GCTA betas were apart from gold standards (multiple regression joint SNP effects). The above comparisons suggest that HAPRAP performs slightly better than GCTA when the individual-level genotypes are available for the entire cohort.

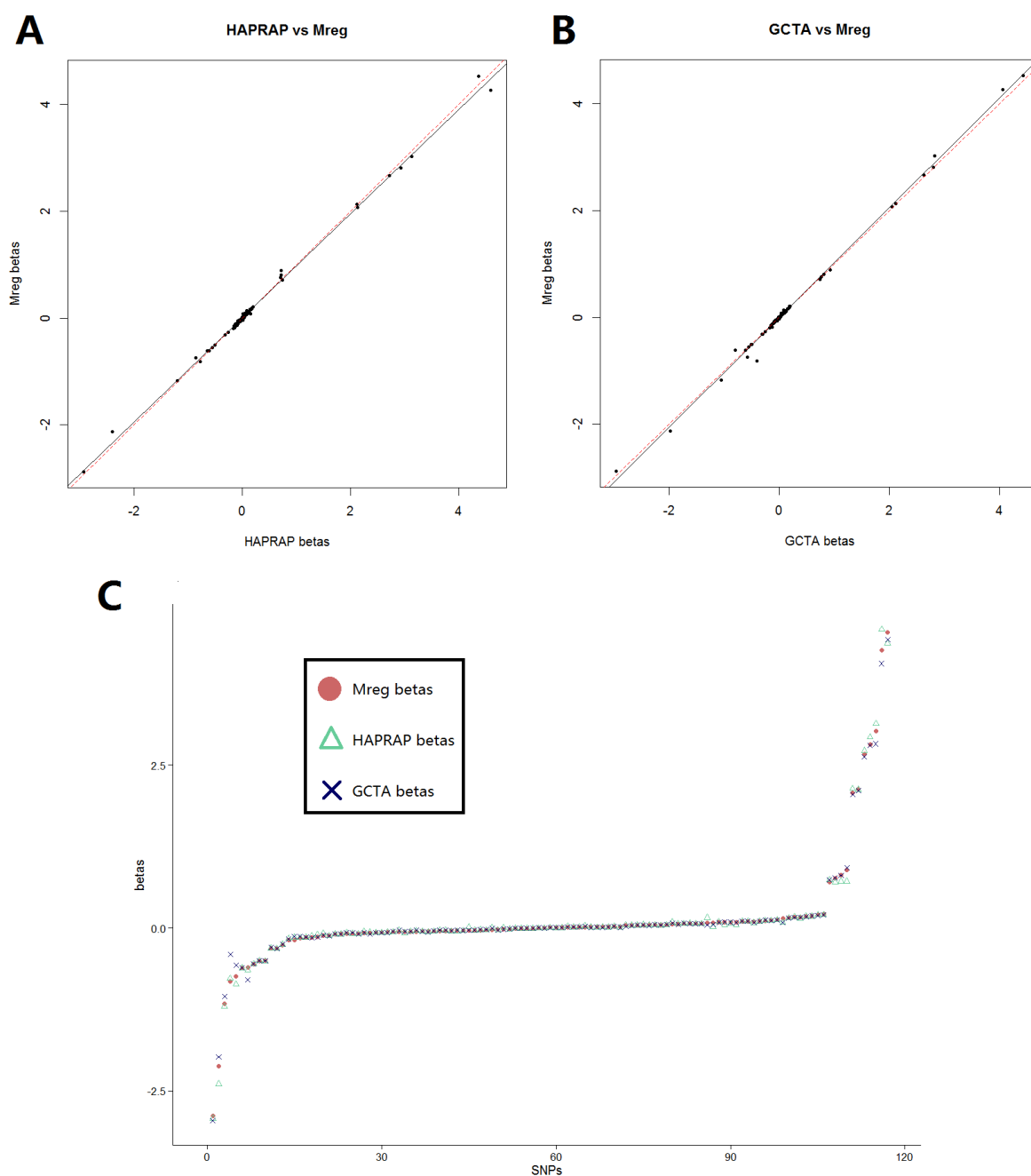


Figure 4-12. Performance Comparison of HAPRAP and GCTA Joint Effect Analysis Using BWHHS Unphased Genotypes.

I drew scatter plots and pair plots to compare the differences between multiple regression (noted as Mreg), joint SNP effects and HAPRAP / GCTA betas for all 117 SNPs inside the region listed in Table 8-9. The entire BWHHS individual-level genotypes and phenotypes were used for the comparison. (A) Scatter plot of Mreg betas vs GCTA betas. (B) Scatter plot

of Mreg betas vs HAPRAP betas. The red line in (A) and (B) is the Y=X line. (C) Pair plot of Mreg betas, HAPRAP betas and Mreg betas. SNPs were sorted by Mreg betas in ascending order. The full circles represent the Mreg betas, clear triangles represent the HAPRAP betas and cross represent the GCTA betas. Detailed results were listed in Table 8-9.

Methods	Cohorts	CCC	95% CI of CCC	C.b
HAPRAP vs Mreg	BWHHS	0.9982	0.9975~0.9987	0.9997
GCTA vs Mreg	BWHHS	0.9978	0.9971~0.9984	0.9996
HAPRAP vs Mreg	1000G	0.9922	0.9887~0.9946	0.9964
GCTA vs Mreg	1000G	0.6129	0.5159~0.6944	0.7562

Table 4-5. *The Results of Concordance Correlation Coefficient Analysis.* Notations: CCC represents the centre estimate of concordance correlation coefficient and 95% CI of CCC is the 95% confidence interval of the centre estimate. Closer the CCC estimates the better agreement of two methods. C.b refers to the bias correction factor, which is a measurement of how far the best-fit line deviates from a line at 45 degrees. No deviation from the 45 degree line occurs when C.b = 1.

4.3.3. A performance comparison of HAPRAP versus GCTA using 1000 Genome genotype data

After further analysing the performance of HAPRAP and GCTA using BWHHS summary statistics and public available genotype data from 1000 Genome, I found that the t-statistics for both methods were normally distributed with means near zero. But the t-statistics deviated substantially this time (Figure 4-13 A and B). After I converted t-statistics to t test p values, the most deviated p value for HAPRAP was 0.412, which suggests that there are no

statistical differences between HAPRAP results and multiple regression results. In contrast, the most deviated t test p value for GCTA was 3.15×10^{-8} , which suggests a significant difference between GCTA beta and multiple regression joint SNP effect of this SNP (Table 8-10). So the GCTA results are not reliable in some cases when using 1000 Genome genotypes as reference panel.

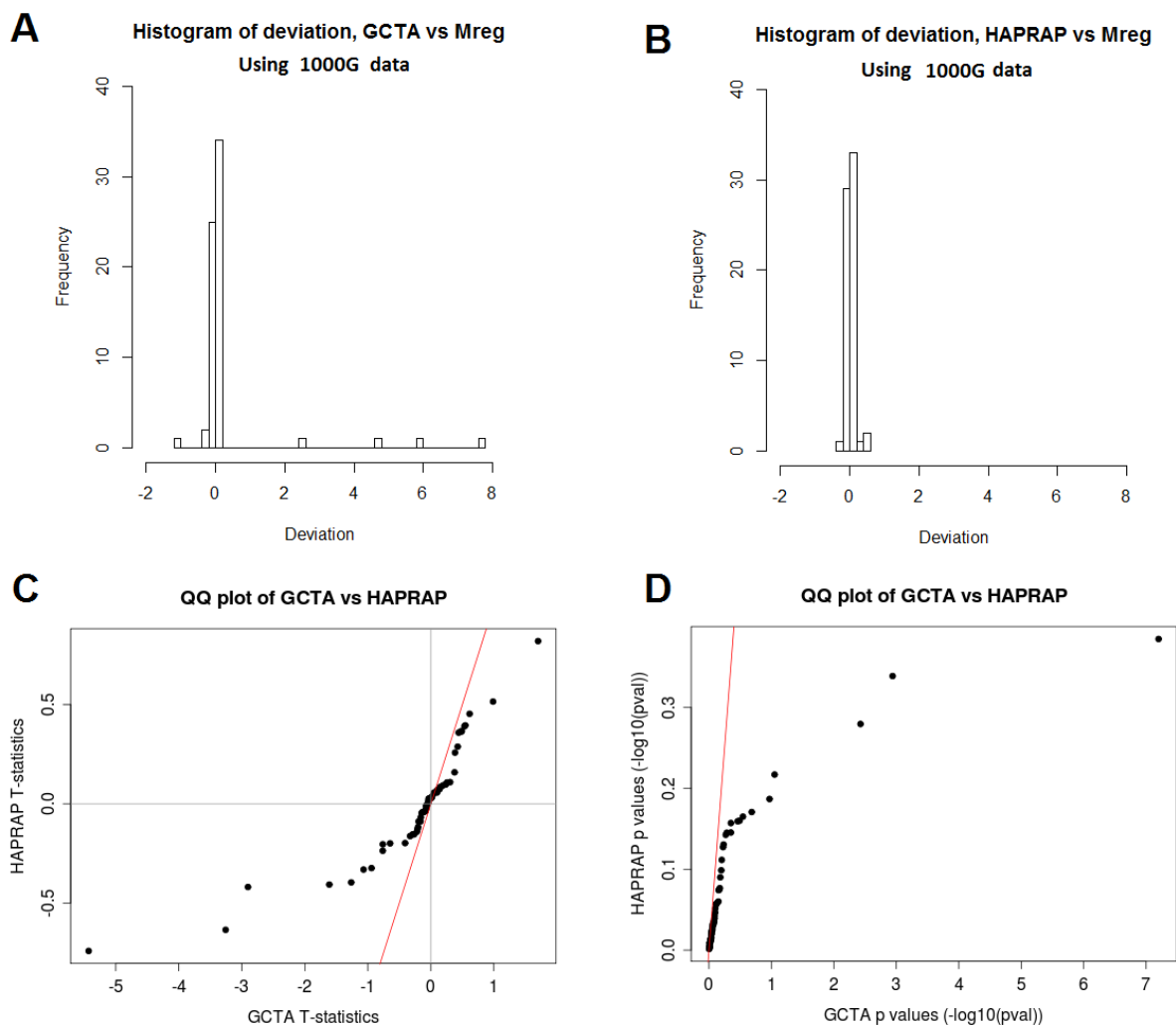


Figure 4-13. Performance comparison between HAPRAP and GCTA joint effect analysis using 1000 Genome phased genotypes. The entire BWHHS summary statistics and 1000 Genome phased individual-level genotypes of 174 people (CEU and GBP) were used for the comparison. The histogram of T-statistics for GCTA and HAPRAP are normal distributed.

From **C** and **D**, some of the SNPs are apart from the $X=Y$ line (red line), where HAPRAP got much larger t-test p values than HAPRAP. This suggests that when a small number of phased individual-level genotypes are available, HAPRAP performance better than GCTA. Detailed results were listed in Table 8-10.

Furthermore, I found that the line of best fit for HAPRAP against multiple regression was again close to the $Y=X$ line (Figure 4-14 A). In contrast, the line of best fit for GCTA against multiple regression had a large angle to the $Y=X$ line (Figure 4-14 B). As the pair plots showed in Figure 4-13 C (and results in Table 8-10), HAPRAP betas were closer to multiple regression joint SNP effects than GCTA betas were. As represented in Table 4-5, CCC analysis results also suggest that HAPRAP got a much better agreement to multiple regression than GCTA (CCC estimate 0.9922 for HAPRAP and 0.6129 for GCTA). The bias correction factor of HAPRAP is much closer to 1 than that of GCTA as well (bias correction factor 0.9964 for HAPRAP and 0.7562 for GCTA). I then estimated the mean square error of the three methods (deviation between points and $Y=X$ line). As shown in Table 8-10, HAPRAP's mean square error (phased by SHAPEIT) is 10.84% lower than that of GCTA-COJO.

These comparisons suggest that HAPRAP is much more reliable than GCTA when utilizing the 1000 Genome individual-level genotypes together with meta-analysis summary statistics.

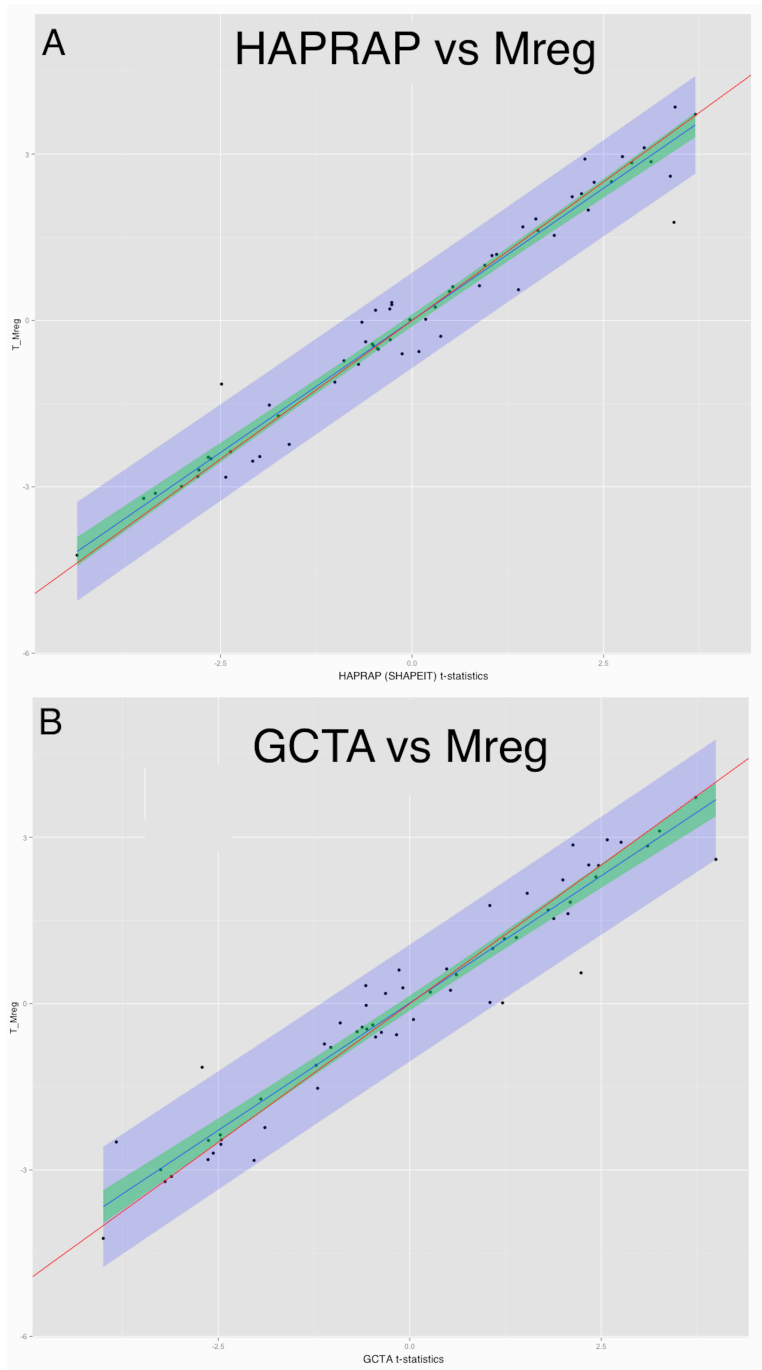


Figure 4-14. Performance Comparison Against HAPRAP and GCTA Joint Effect Analysis Using 1000 Genome Phased Genotypes. (A) Scatter plot of Mreg t-statistics vs HAPRAP (SHAPEIT) t-statistics. (B) Scatter plot of Mreg t-statistics vs GCTA-COJO t-statistics. In (A) and (B), the longer light line is the $Y=X$ line, the shorter dark line is the line of best fit. Inner zone and outer zone are the 95% confidence interval and prediction interval of the line of best fit

respectively. The full circles represent the Mreg betas, clear triangles represent the HAPRAP betas and cross represent the GCTA betas. Detailed results were listed in Table 8-10.

4.4. Discussion

In this chapter, I presented an EM style method, HAPRAP, which provides a new angle to perform regional association analysis using meta-analysis results. This method is built on the underlying biology of haplotypes, which lead to 3 significant advantages against GCTA-COJO: It is more accurate when LD measure r^2 is high (i.e. $r^2 > 0.8$) and sample size in the reference panel is low (i.e. $N < 500$). It is more accurate when analysing rare variants (i.e. $MAF < 0.01$) than methods using pair-wise LD; low MAF presents no barrier to performance. Haplotype-based analyses lead to identification of the most informative SNP. The independent association SNP need not be the most strongly associated SNP within the genomic region.

4.4.1. Advantages of HAPRAP

From our independent benchmark, we proved that the performance of HAPRAP is comparable to multiple regression when all the individual-level phenotypes and genotypes are available.

As mentioned in the introduction, most of the individual-level data will be not available when conducting a meta-analysis of published consortial data. We have proved in the independent benchmark that applying multiple regression to the available individual-level data, in a single accessible cohort, will return much less accurate results than using HAPRAP and GCTA-which utilize the same individual-level genotypes from one cohort but also make

use of the overall meta-analysis summary statistics. My performance comparison using the 1000 Genomes Project data showed that HAPRAP's mean square error was on average 10% lower than that of GCTA-COJO. Furthermore, the major difference between the two methods is that HAPRAP utilizes phased haplotypes whereas GCTA utilizes unphased genotypes.

As I have mentioned in the introduction section, haplotypes contain more information than unphased genotypes. Thus, HAPRAP should have a theoretical advantage over GCTA for analysing individuals with accurate phased haplotypes. For instance, the 1000 Genome consortium (1000 Genomes Project Consortium et al., 2010) used Segmented HAPlotype Estimation and Imputation Tool (SHAPEIT) (Delaneau et al., 2013) to phase the genotypes with family information. I proved this advantage by using the admixture European population from the 1000 Genome data. Since the sample size of 1000 Genome data is still small (around 200 individuals for each major ethnic group), the risk of using it as a reference panel is relatively high. The errors of using the 1000 Genomes Project data as a reference panel are relatively high since the sample size is currently small (under the null, variance of LD correlation is $1/n$ with n being the sample size. Given $n=200$, the standard error of LD correlation estimate is ~ 0.07). However, for researchers not involved in the meta-analysis or those yet to use the data, it is still meaningful to use 1000 Genome data as a test dataset to get a general pattern for the multiple potentially associated SNPs in the region(s) of interest. With the development of open access genotype databases such as 1000 Genomes Project and HAPMAP phase 3 (International HapMap 3 Consortium et al., 2010), more and more free data will be available. HAPRAP's accuracy advantage against GCTA will increase.

In addition, GCTA adjusts genotypes based on MAFs of SNPs. The genotype for a certain SNP under the additive model will be $0-2p$, $1-2p$ and $2-2p$, where p is the reference allele

frequency of this SNP. This adjustment is due to the assumption that SNPs with lower MAF will have higher effects (Park et al., 2011), which will mislead the regression if this is the case. In contrast, HAPRAP uses haplotype frequencies instead of allele frequencies, which does not rely on the above assumption. This suggests that HAPRAP may have an advantage over GCTA for rare variants. The performance comparison of rarer variants also suggest that HAPRAP may have advantage against GCTA when $MAF < 0.1$. Moreover, we highlighted a rare variant in Apolipoprotein B (APOB), rs41288783, as an example in real world (Table S2). The SNP had a MAF of 0.0018 in BWHHS individuals. The joint SNP effect of this SNP estimated by HAPRAP is near to the gold standard comparing to the effect estimated by GCTA-COJO (gold standard: 0.731, HAPRAP: 0.705, GCTA-COJO: 0.449).

4.4.2. Assumptions and limitations for HAPRAP

There is an uncertainty when phasing haplotypes from genotypes. The haplotype phasing approach I selected is an E-M algorithm. It is not accurate especially when there are more than 10 SNPs in the model (Browning and Browning, 2011). In fact, in the same paper, Browning SR et al illustrate that the phasing accuracy can be increased up to 98% by using hidden Markov model (HMM) haplotype phasing tools such as BEAGLE (Browning and Browning, 2009), IMPUTE2 (Howie et al., 2009) and MACH (Li et al., 2010). However, these tools are computationally expensive. I have opted to use empirical E-M algorithm here for two reasons. Firstly, the method is computationally fast. Secondly, it is almost impossible to identify 10 or more multiple associated SNPs in a specific genomic region and the errors caused by haplotype uncertainty are a minor issue in the step-wise elimination process.

Moreover, if SNPs with a very high variance inflation factor (VIF) are included, HAPRAP will return extremely large betas for a pair of SNPs. So it is necessary to remove SNPs with VIF

higher than 10 step-wisely before applying HAPRAP.

HAPRAP needs more time to finish the analysis than GCTA. The reasons for this are listed below. Firstly, HAPRAP needs time to phase haplotypes if the genotypes are unphased; secondly, it is time consuming to conduct the standard errors of the independent SNP effects using simHAPRAP. To improve the efficiency of HAPRAP, I developed the pre-significant test system to exclude insignificant SNPs from the model so that the only SNPs within the optimal model will be conducted. For the 6 SNPs model of *ABCG5-8* regions (Chapter 6), HAPRAP finished the pre-significant test within only 10 seconds but took another 622 seconds (approximately 10 minutes) to calculate the standard errors of the remaining SNPs using simHAPRAP. I think it is still reasonable to use 10 minutes for a fine mapping analysis of a genomic region.

Normally, two SNPs are unlikely to be in a strong LD if they are far away from each other. However, long-distance LD may occur during an extreme reduction in population size (note as a population bottleneck) (Zhang et al., 2004). For example, long-distance LD had been found over a few hundred kb in unrelated population (Sabeti et al., 2002). For convenience of analysis and simplicity of SNP models, I set a threshold of genomic regions of 1MB for HAPRAP so that it can process regional association analysis more efficiently. In some specific cases, two SNPs can be in strong LD even if they are more than 1MB to each other (Yang J et al, 2012). So I suggest checking long distance LD inside each chromosome after the optimal SNP models were identified by HAPRAP.

4.4.3. Conclusion

In this chapter, I have described the novel fine mapping method HAPRAP which I developed. This method uses haplotypes instead of genotypes to represent the LD correlation among

multiple SNPs. The method can be applied to both quantitative traits and disease outcomes. Compared to other related methods, the results of HAPRAP are highly consistent with multiple regression results using individual-level data. Moreover, HAPRAP consistently outperforms GCTA joint SNP effect analysis at group-level meta-analysis across different linkage disequilibrium (LD) correlations (r^2) and sample sizes of reference panels. The E-M algorithm I have developed is computational fast but the standard error simulation needs time but still acceptable. This method is developed for region association analysis for SNPs within 1Mb.

CHAPTER 5. FINE MAPPING OF META-ANALYSIS GROUP-LEVEL DATA USING HAPRAP

5.1. Introduction

In Chapter 3, I have introduced the cardiovascular disease and the genetic variants associated with some CVD related phenotypes such as ECG traits and GBD traits. Other traits such as blood lipids traits and coagulation traits are also related to CVD risks. Identifying the genetic variants associated with these traits and interpreting the causality are of equal importance.

5.1.1. Blood lipids traits

Plasma lipids are mostly transported in a protein capsule. The major content of blood lipids are cholesterol and fatty acids. Cholesterol is an organic molecule which is an essential structural component of animal cell membranes that is required to establish proper membrane permeability and fluidity. Cholesterol also works as a precursor for the synthesis of steroid hormones, bile acids, and vitamin D (Hanukoglu, 1992). The liver cells normally produces more cholesterol than other cells in humans.

Fatty acids are a kind of carboxylic acid with a long aliphatic tail. The tail can be either saturated or unsaturated. Most naturally synthesizing fatty acids have a chain of an even number of carbon atoms, from 4 to 28. Two important source of fatty acids are triglycerides and phospholipids. Fatty acids are important sources of energy for human beings since they

can yield large quantities of Adenosine triphosphate (ATP) when they are metabolized. ATP is the most important element for intracellular energy transfer (Knowles, 1980). Many cell types can accept fatty acids as a source of energy, particularly, heart and skeletal muscle prefer fatty acids. The brain can use fatty acids as well (Marin-Valencia et al., 2013)

Since lipids are not absorbed in water (such as bloodstream), lipoprotein molecules enable the transportation of cholesterol and triglycerides. In order of molecular size, lipoprotein can be categorized into chylomicrons, very low-density lipoprotein (VLDL), intermediate-density lipoprotein (IDL), low-density lipoprotein (LDL) and high-density lipoprotein (HDL). The pathway of cholesterol and lipoprotein transport are shown in Figure 5-1.

heritable risk factors for cardiovascular disease and targets for therapeutic intervention (Kathiresan et al., 2007). Environmental factors such as smoking, diet and physical activity all play a role in determining individual lipid profiles. Moreover, previous family studies suggest that genetic constitutions may contribute to half of the variation in these traits in many populations (Pilia et al., 2006) (Pollin et al., 2004). A previous GWAS study identified 95 loci associated with blood lipids, accounting for 10%~12% of total trait variance (Teslovich et al., 2010). Recently meta-analysis of even more GWASs combining European and non-European-ancestry individuals identified 157 loci associated with lipid levels at $P < 5 \times 10^{-8}$ (Global Lipids Genetics Consortium et al., 2013).

HDL cholesterol

Ranked by size, HDL is the smallest particles of the lipoprotein. HDL particles move fat acids and cholesterol out from cells, including within artery wall atheroma and transport it back to the liver for excretion or re-utilization. So the cholesterol carried within HDL particles (HDL-C) is sometimes called "good cholesterol". A person with higher levels of HDL-C tend to have a better cardiovascular health (Sirtori and Fumagalli, 2006), while those with low HDL-C cholesterol levels (especially less than 40 mg/dL or about 1 mmol/L) have increased rates for heart disease (Toth, 2005). Epidemiological studies have highlighted the antiatherogenic function of HDLC and showed that an increase of 1 mg/dL in HDLC levels is associated with a 2% and 3% decrease in risk of coronary artery disease in men and women, respectively.

As shown in Table 8-11, previous GWASs have identified 107 SNPs associated with HDL-C. Some well-known gene such as Cholesteryl ester transfer protein (*CETP*), lipase hepatic (*LIPC*) and lipoprotein lipase (*LPL*) were fine mapped. *CETP*, which located on chromosome 16q21,

transfers cholesteryl esters between lipoproteins. *CETP* could be targeted in *CETP* deficient individuals to increase HDL-C levels (Agerholm-Larsen et al., 2000). To be able to identify suitable targets, it will be necessary to fine map these genomic regions. Multiple associated SNPs were identified using both European-descent based and African-descent based meta-analysis (Teslovich et al., 2010) (Buyske et al., 2012). LPL encodes lipoprotein lipase. It is expressed in cell types such as heart, muscle, and adipose tissue. LPL functions as a homodimer, and has the dual functions of triglyceride hydrolase and ligand/bridging factor for receptor-mediated lipoprotein uptake. *LIPC*, which encodes hepatic triglyceride lipase, is an important enzyme in HDL metabolism. It is mostly expressed in liver.

LDL cholesterol

LDL molecules can transport their content of many fat molecules into artery walls, attract macrophages, and thus drive atherosclerosis. So they are often noted as bad cholesterol for human beings. High concentrations of LDL cholesterol are associated with increased risk of coronary artery disease (CAD) and the leading cause of death in United States (Roger et al., 2012). LDL-C level is influenced by the environment factors and genetic factors. Among the total variation, approximately 40%–50% of them is heritable (de Miranda Chagas et al., 2011).

As shown in Table 8-12, previous GWASs have identified 73 loci associated with LDL, which only explained 10%-12% of the total variance in LDL-C (Teslovich et al., 2010).

A recently fine mapping study of five LDL cholesterol associated loci claimed that variants with the strongest association were often substantially different in effect allele frequencies and effect sizes from those identified by GWAS (Sanna et al., 2011). To explain the missing heritability and to replicate the GWASs results, fine mapping is needed for candidate gene

regions such as Apolipoprotein E (*APOE*), Poliovirus Receptor Related 2 (*PVRL2*), Translocase Of Outer Mitochondrial Membrane 40 Homology (*TOMM40*), Low-Density Lipoprotein Receptor (*LDLR*), Cadherin, EGF LAG Seven-Pass G-Type Receptor 2 (*CELSR2*) - Proline/Serine-Rich Coiled-Coil 1 (*PSRC1*) region and Apolipoprotein B (*APOB*).

APOE is a gene with three major isoforms: *ApoE2* (cys112, cys158), *ApoE3* (cys112, arg158), and *ApoE4* (arg112, arg158) (Ghebranious et al., 2005). These isoforms are defined by the combinations of two SNPs rs429358 and rs7412 (Table 5-1).

	rs429358	rs7412
E2	T	T
E3	T	C
E4	C	C

Table 5-1. *Three major isoforms of APOE gene*

E2 can be found in approximately 7% of the population (<http://www.alzgene.org/meta.asp?geneID=83>). E2 is associated with increased risk for atherosclerosis and Parkinson's disease (Breslow et al., 1982) (Feussner et al., 1998) (Civeira et al., 1996) (Huang et al., 2004). For E3, it is the most common *APOE* genotype which can be found in approximately 79% of the population (<http://www.alzgene.org/meta.asp?geneID=83>). More important, E4 has been associated with disorders such as atherosclerosis (Federoff et al., 2012), Alzheimer's disease (Mahley, 1988) (Corder et al., 1993), impaired cognitive function (Strittmatter et al., 1993) (Deary et al., 2002). Its prevalence is about 14% of the population (<http://www.alzgene.org/meta.asp?geneID=83>).

LDLR is a protein coding gene located in chromosome 19p3.2. It binds LDL and transport it into cells by endocytosis. Mutation in *LDLR* is also associated with hypercholesterolemia (<http://www.genecards.org/cgi-bin/carddisp.pl?gene=LDLR>).

In addition, *APOB* is a protein-coding gene of Apolipoprotein B, which is a major protein constituent of chylomicrons, LDL and VLDL (<http://www.genecards.org/cgi-bin/carddisp.pl?gene=APOB>).

Triglycerides (TG)

TG is an ester derived from glycerol and three fatty acids. As a part of the plasma lipid, it helps enable the double sided transference of adipose fat and blood glucose from the liver.

As other blood lipids, TG level is heritable risk factors for cardiovascular disease (Namboodiri et al., 1985) (KANDEL et al., 1961) (Miller and Miller, 1975).

As shown in Table 8-13, totally 79 SNPs were found to be associated with TG from previous GWASs. I will fine mapping candidate genes such as Zinc Finger Protein 259 (*ZNF259*), Apolipoprotein A-V (*APOA5*), BUD13 Homolog (*BUD13*), LINC00900 long intergenic non-protein coding RNA 900 (*LOC283143*) and *LPL* (which is associated with HDL-C level as well) in this chapter. For example, the common *LPL* nonsense mutation S477X (rs328) in this region is known to be associated with TG. And multiple associated signals were found by the previous GWASs (Table 8-13).

Total cholesterol (TC)

Until now, 62 SNPs were proved to be associated with TC which explained 12.4% of the total heritability (Global Lipids Genetics Consortium, 2013). Table 8-14 listed the most recent 15

SNPs found to be associated with TC. Candidate genes such as *APOE*, *LDLR*, *PVRL2* and *APOB* need further fine mapping.

5.1.2. Coagulation

Plasma levels of coagulation factors such as VII (FVII), VIII (FVIII) and von Willebrand factor (vWF) is the basis of hemostasis, which prevents life-threatening blood loss from damaged blood vessels. Both coagulation factors VII and VIII are produced in liver and play a key role in the initiation and propagation of fibrin formation (Smith et al., 2010b). Increased circulating levels of FVIII and vWF will increase risk of several thrombosis (Koster et al., 1995) (Folsom, 2001). Deficiency in FVII and vWF are causality of Hemorrhagic complications (Cooper et al., 1997) (Acharya et al., 2004). X-linked deficiency in FVIII, which is called haemophilia A clinically, is associated with hemorrhagic complications as well (Antonarakis et al., 1987). Plasma levels of these factors will be influenced by both environmental and genetic factors (Conlan et al., 1993) (de Lange et al., 2001).

Factor VII

In the tissue-factor pathway, activated FVII is a catalyst for factor X (FX) activation, which converts prothrombin to thrombin. The heritability estimate of FVII is from 0.53 to 0.63 (Smith NL et al, 2010). Table 8-15 shown 4 SNPs associated with FVII from previous GWASs.

Factor VIII

During the propagation process, activated FVIII can activates FX when the activated factor IX present. The heritability estimate of FVIII is from 0.4 to 0.61 (Smith et al., 2010b). Table 8-16 shown 11 SNPs associated with FVII from previous GWASs.

Von Willebrand (vWF)

Von Willebrand factor is produced by endothelial cells and megakaryocytes, and plays multiple roles in hemostasis. Firstly, it acts as an adhesion molecule that anchors platelets to exposed collagen after endothelial cell damage. Secondly, serves as a carrier protein for FVIII, therefore extending the half-life of FVIII. The heritability estimate of vWF is from 0.31 to 0.75 (Smith et al., 2010b). Table 8-17 shown 11 SNPs associated with vWF from previous GWASs.

To sum up, there are only a limited number of SNPs reported to be associated with coagulation factors. I will fine map the following genomic regions using HAPRAP and GCTA: *F7*, *F10*, *PROCR*, *TRPC4AR* and *MYH7B* for FVII; *ABO*, *OBP2B*, *SURF4*, *C9orf96*, *ADAMTS13*, *CACFD1* and *GBGT1* for FVIII; *ABO*, *OBP2B*, *SURF4*, *C9orf96*, *ADAMTS13*, *CACFD1*, *MED22*, *SLC2A6* and *GBGT1* for vWF.

5.1.3. Other CVD related traits

C-reactive protein (CRP)

C-reactive protein, which is found in blood, is a pattern-recognition molecule of innate immunity. Low level of plasma concentrations of CRP is an independent predictor of diseases such as incident metabolic syndrome, type 2 diabetes, myocardial infarction, and stroke in both men and women (Ridker et al., 1997) (Pradhan et al., 2001) (Laaksonen et al., 2004) (Ridker et al., 2002) (Han et al., 2002). CRP levels will be affected by both environmental factors, such as obesity, smoking and hormone-replacement therapy, and genetic factors (Hage and Szalai, 2007). Previous GWASs have provided insight into mechanisms underlying relationships between CRP, insulin resistance, metabolic syndrome,

and vascular events (Table 8-18). Further analysis is need for candidate genes such as *APOE*, *CRP*, *NHF1A* and *C12orf43*.

Interleukin 6 (IL6)

Interleukin 6 encodes a cytokine that works in inflammation and maturation of B cells. So its function is associated with several inflammation diseases such as diabetes mellitus and systemic juvenile rheumatoid arthritis
(<http://www.genecards.org/cgi-bin/carddisp.pl?gene=IL6>).

Bilirubin

Bilirubin is a major component of bile and it is the breakdown product of heme catabolism. Increased levels of serum bilirubin is associated diseases such as gallstone disease. Moreover, the inverse correlation between bilirubin levels and atherosclerosis risk had been found and supported by meta-analysis (Hunt et al., 2001) (Djoussé et al., 2001) (Novotný and Vitek, 2003), which suggests that it can be protective.

The heritability of bilirubin levels had been identified by family-based studies (Kronenberg et al., 2002) (Lin et al., 2003) (Clementi et al., 2007). Furthermore, linkage analyses identified the *UGT* region on chromosome 2q37.1 as the major genomic region for bilirubin levels (Kronenberg et al., 2002) (Lin et al., 2003).

Pervious GWASs found 38 SNPs associated bilirubin, which is listed in Table 8-19. *UGT* and *USP40* will be extremely interesting regions to follow-up.

Serum urate (surate)

Uric acid is a heterocyclic compound with the formula $C_5H_4N_4O_3$. It is the end product of purine metabolism in humans. Urate is known as the salts forms from uric acid.

Approximately 70% of urate disposal occurs via the kidneys. Impaired renal excretion leads to hyperuricemia in 5~25% of the human population (Becker and Jolly, 2006).

Hyperuricemia will lead to development of gout, which is an inflammatory arthritis, in about 10% of the population.

Until now, 68 SNPs were identified to be associated with urate by previous GWASs (Table 8-20). SLC2A9 gene, which is the first gene to be found by GWASs, is a known fructose transporter (Manolescu et al., 2007).

5.1.4. Atopic Dermatitis

Atopic dermatitis, also called eczema, is a common chronic inflammatory skin diseases with prevalence of 20% in children and 3% in adults. It is commonly starts during infancy. Food allergy, asthma and rhinitis are often occurs in eczema patients. Its clinical characters includes dry skin, intense pruritus and age-related inflammatory lesions (Bieber, 2008). Key components affecting the development of atopic dermatitis include change in skin barrier function and immunologic abnormalities. Further research is needed to identify the exact molecular mechanisms of atopic dermatitis.

Atopic dermatitis is strongly inherited. Until now, 28 SNPs were reported to be associated with atopic dermatitis (Table 5-21). Specific genomic regions such as FLG in chromosome 1, *C11orf30* in chromosome 11 need further investigation.

5.1.5. Aims and Objectives

In the above section, I have listed the traits and their associated genomic regions I am interested in. I would conduct fine mapping to these candidate genes using meta-analysis summary results directly. In Chapter 5, I have presented the novel fine mapping method HAPRAP and compare its performance with other methods. In this chapter, I will focus on the application of HAPRAP and GCTA joint SNP effect analysis to these traits. Secondary association signals will be identified using meta-analysis summary results directly.

Objectives

- Further fine mapping the ECG traits and GBD traits
- Fine mapping the four major blood lipids traits
- Fine mapping coagulation traits
- Fine mapping other CVD related traits including
- Fine mapping another binary traits, atopic dermatitis.
- Compare the difference of meta-analysis results between using fix effect model and random effect model

5.2. Methods and Materials

5.2.1. Sample datasets

BWHHS individual-level data

BWHHS individual-level genotypes of 1980 elder ladies from 60-79 years old were used in this chapter as the reference genotype panel. The individual-level genotypes were genotyped using the Metabochip chip for these ladies. The BWHHS cohort was pre-phased using MACH (Li et al., 2010) and imputed using the 1000 Genomes CEU reference panel on the imputation tool minimac (Howie et al., 2012).

UCLEB Meta-analysis of CVD related traits

I used summary statistics of UCLEB meta-analysis on totally 14959 individuals.

The UCLEB meta-analysis of ECG traits included three participating studies on totally 7106 individuals: the BWHHS (Lawlor et al., 2003), the British Regional Heart Study (BRHS) (Shaper et al., 1981) and the Whitehall II study (WHII) (Marmot et al., 1991).

The UCLEB meta-analysis of blood lipids traits included eight participating studies on totally 14959 individuals: MRC NSHD (Wadsworth et al., 2006), ELSA (Marmot et al., 2003), ET2DS (Price et al., 2008), EAS (Fowkes et al., 1991), CAPS (Bainton et al., 1992), BWHHS, BRHS and WHII.

The UCLEB meta-analysis of vWF and FVII included five participating studies on totally 9009 individuals: EAS, CAPS, BWHHS, BRHS and WHII. Besides, FVIII meta-analysis included four independent studies on totally 5872 individuals: EAS, CAPS, BWHHS and BRHS.

For other traits, CRP included seven studies on totally 11545 individuals: ELSA, ET2DS, EAS, CAPS, BWHHS, BRHS and WHII. IL6 included six studies on totally 9541 individuals: ET2DS, EAS, CAPS, BWHHS, BRHS and WHII. Bilirubin included four studies on totally 6530 individuals: ET2DS, CAPS, BWHHS and BRHS. Serum urate included two studies on 4242 individuals: BWHHS, BRHS

Meta-analysis of GBD

I used meta-analysis summary statistics of a consortium analysis of gallbladder disease (GBD) (Rodriguez et al., 2014). The meta-analysis included three individual studies from both UK and USA: the Atherosclerosis Risk in Communities Study (ARIC) (The ARIC investigators, 1989), the Whitehall I (WHI) (The Women's Health Initiative Study Group, 1998) and the BWHHS. Totally 15213 individuals were included in this meta-analysis.

EAGLE consortium meta-analysis of atopic dermatitis

The EAGLE consortium meta-analysis of atopic dermatitis was used in this section (Paternoster et al., 2012). Totally 119496 European individuals were involved in the meta-analysis. Details of individual studies involved in this meta-analysis were listed in Chapter 2.1.7).

5.2.2. Fine mapping for GBD outcomes

HAPRAP is appropriate to case-control studies. I demonstrated this by using meta-analysis summary statistics of the consortium analysis of gallbladder disease (GBD) and individual-level genotypes using the HumanCVD chip for 3078 BWHHS individuals. I selected two head to head genes, *ABCG5* and *ABCG8*, as an example. This region has been previously

shown to contain multiple associated SNPs with GBD (Buch et al., 2007). Based on the summary statistics from the meta-analysis, eight SNPs were significantly associated with GBD in this region. Two of them were removed because of high VIF. I applied HAPRAP and GCTA to the remaining six SNPs with meta-analysis p values smaller than 5×10^{-5} (Table 8-22).

5.2.3. Fine mapping for UCLEB meta-analysis of CVD related traits

I applied HAPRAP using summary statistics from the UCLEB meta-analysis on totally 13 different traits. I used 1980 BWHHS individuals of recent European ancestry, which is genotyped using MetaboChip, as a reference genotype panel. As I have mentioned above, BWHHS cohort is part of the UCLEB meta-analysis. The meta-analysis significant p value cut-off is 1×10^{-5} . For quality control of SNPs, I used the most likely genotypes of the imputed SNPs and excluded SNPs with MAF < 0.05. SNPs with VIF < 10 were selected for further fine mapping.

For ECG traits, I focused on the *NOS1AP* region, which has been previously reported to be associated with QTc interval and *SCN5A-SCN10A* region associated with PR interval.

For plasma lipids traits, I selected three regions associated with HDL-C: *CETP-HERPUD1*, *LIPC* and *LPL* associated with HDL; seven regions associated with LDL-C: *APOE*, *PVRL2*, *TOMM40*, *LDLR-SMARCA4*, *CELSR2*, *APOB* and *PCSK9*; five regions associated with TG: *ZNF259*, *APOA5*, *BUD13*, *LOC283143* and *LPL*; four regions associated with TC: *APOE*, *LDLR-SMARCA4*, *PVRL2* and *APOB*.

For coagulation traits, I chose five regions associated with FVII: *F7*, *F10*, *PROCR*, *TRPC4AP* and *MYH7B*; seven regions associated with FVIII: *ABO*, *OBP2B*, *SURF4*, *C9orf96*, *ADAMTS13*,

CACFD1 and *GBGT1*; nine regions associated with vWF: *ABO*, *OBP2B*, *SURF4*, *C9orf96*, *ADAMTS13*, *CACFD1*, *MED22*, *SLC2A6* and *GBGT1*.

For other traits, I selected *IL6* gene associated with IL6; *SLC2A9* gene associated with sUrate; four genes associated with CRP: *APOE*, *CRP*, *HNF1A* and *C12orf43*; three regions associated with bilirubin: *UGT*, *USP40* and *SLC22A3*.

Totally 394 meta-analysis significant SNPs passed the quality control process for the above regions. The SNPs, their genomic regions and their relative traits were listed in Table 8-23.

For all the above traits, the p value threshold of the fine mapping analysis was 5×10^{-5} .

5.2.4. Fine mapping for Atopic Dermatitis

I applied HAPRAP to summary statistics from the EAGLE genome-wide association meta-analysis of atopic dermatitis, made available to me by the EAGLE consortium. The meta-analysis was conducted using European only individuals and a fixed effect model. All SNPs reach the meta-analysis significance of p value $< 5 \times 10^{-8}$ were selected for the further fine mapping analysis.

9233 ALSPAC individuals were use as the genotype reference panel for HAPRAP. The EAGLE consortium aim to set more accurate conditional analysis p value threshold in each candidate region. To achieve that, the EAGLE consortium used the SNP spectral decomposition (SNPSpD) method to calculate the number of effective tests in each region (Nyholt, 2004) and the conditional analysis p -value threshold for each region were then calculated based on the number of effective tests. The p -value threshold provided by the EAGLE consortium were listed in Table 5-2.

Chr	P value threshold	Start region	End region
1	6.45E-06	150250000	154500000
2	9.29E-05	102777103	103277103
3	9.85E-05	112180000	112641174
3	9.82E-05	52838790	53338790
4	1.30E-04	122993592	123493592
5	0.00012	131741085	132150000
6	6.31E-06	32325524	32825524
6	1.14E-05	31216217	31716217
8	1.60E-04	81090000	81340000
10	7.34E-05	6020000	6288853
10	8.03E-05	64119999	64619999

11	9.26E-05	65309266	65809266
11	7.52E-05	76031593	76531593
11	2.20E-04	128040000	128190000
14	9.32E-05	35309126	35809126
16	6.27E-05	10979589	11479589
17	1.00E-04	40278131	40778131
19	6.91E-05	8539722	9039722
20	5.24E-05	62053115	62553115

Table 5-2. *P value threshold for each atopic dermatitis candidate region.*

5.3. Results

5.3.1. Fine mapping of GBD outcomes

I conducted the fine mapping analysis of loci associated with GBD using both HAPRAP and GCTA. Both methods identified rs4299376 and rs4953023 as independent effect SNPs (Table 5-3). Besides, both methods return similar joint SNP effects and standard errors, which proved the suitability of HAPRAP in the case-control study.

SNP	CHR	POS	MAF	HAPRAP			GCTA		
				BETA	SE	p value	BETA	SE	p value
rs4299376	2	43926080	0.315	-0.223	0.033	8.86×10^{-12}	-0.221	0.033	1.92×10^{-11}
rs4953023	2	43927504	0.066	0.688	0.054	5.87×10^{-38}	0.712	0.053	5.50×10^{-41}

Table 5-3. *HAPRAP and GCTA joint effect analysis of identifying two independent effect SNPs associated with gallbladder disease in ABCG 5 and 8 regions.*

5.3.2. Fine mapping for ECG traits

I processed the fine mapping analysis using UCLEB meta-analysis results of ECG traits. For *NOS1AP* gene associated with QTc interval, I identified three jointly associated SNPs:

rs10429888, rs16857031 and rs10918859 using HAPRAP (Table 5-4). In contrast, GCTA only detected the top hit, rs12143842, as independent.

For *SCN5A-10A* region associated with PR interval, I identified three independent effect SNPs: rs12635898, rs73056438, rs6801957 using HAPRAP. In comparison, GCTA identified the same three SNPs as HAPRAP.

5.3.3. Fine mapping for lipids traits

I performed the fine mapping analysis using UCLEB meta-analysis results of plasma lipids traits. As shown in Table 5-5, totally 41 SNPs in 14 loci have multiple association SNPs for four lipids traits: 5 SNPs in 2 loci associated with HDL-C; 9 SNPs in 4 loci associated with LDL-C; 19 SNPs in 4 loci associated with TG and 8 SNPs in 3 loci associated with TC.

5.3.4. Fine mapping for coagulation traits

I conducted the fine mapping analysis using UCLEB meta-analysis results of coagulation traits. As shown in Table 5-6, totally 50 SNPs in 15 loci have multiple association SNPs for four lipids traits: 2 SNPs in F7 gene associated with FVII; 13 SNPs in 5 loci associated with FVIII and 35 SNPs in 9 loci associated with vWF.

SNP	Chr.	Location (bp)	Nearest gene	Ref Allele	MAF	UCLEB meta-analysis			HAPRAP, haplotypes from BWHHS		
						BETA	Std Err	P	BETA	Std Err	P
rs10429888	1	162024987	<i>NOS1AP</i>	A	0.1291	4.0402	0.6642	1.18x10 ⁻⁹	3.7071	0.642	8.07x10 ⁻⁹
rs16857031	1	162112910	<i>NOS1AP</i>	G	0.1469	2.8277	0.6276	6.61x10 ⁻⁶	2.8386	0.6082	3.104x10 ⁻⁶
rs10918859	1	162169268	<i>NOS1AP</i>	A	0.1784	3.5747	0.5825	8.41x10 ⁻¹⁰	2.5991	0.5553	2.913x10 ⁻⁶
rs12635898	3	38601069	<i>SCN5A-10A</i>	A	0.3909	2.5793	0.4753	5.73x10 ⁻⁸	2.9567	0.4916	1.84x10 ⁻⁹
rs73056438	3	38646478	<i>SCN5A-10A</i>	A	0.3572	2.6866	0.5318	4.38x10 ⁻⁷	3.7213	0.5575	2.66x10 ⁻¹¹
rs6801957	3	38742319	<i>SCN5A-10A</i>	A	0.4526	3.9874	0.4243	5.61x10 ⁻²¹	4.4573	0.4492	4.69x10 ⁻²³

Table 5-4. Summary of multiple associated SNPs at *NOS1AP* for QTc interval and *SCN5A-10A* for PR interval. The UCLEB meta-analysis betas are the marginal SNPs effects of the SNPs. For the HAPRAP analysis, UCLEB meta-analysis summary statistics and individual-level haplotypes from 1980 BWHHS individuals were used. The HAPRAP betas are the independent SNP effects of the SNPs. Chr represents the chromosomes of the SNPs. MAF represent the minor allele frequencies of the SNPs. The Standard errors are noted as Std Err and the P values are noted as P.

Locus	Chr	independent SNP	traits	Allele	Freq	Beta	P	HAPRAP Beta	HAPRAP P	GCTA Beta	GCTA P
<i>CETP</i>	16	rs12708967	HDL	C	0.3016	-0.0617	1.87E-33	-0.0386	5.89E-07	-0.0367	7.25E-12
<i>CETP</i>	16	rs3764261	HDL	A	0.4449	0.0822	7.35E-80	0.0795	1.42E-36	0.0589	1.26E-36
<i>CETP</i>	16	rs5883	HDL	A	0.1264	0.0566	7.49E-10	0.095	7.52E-15	0.0913	1.35E-21
<i>LIPC</i>	15	rs1077834	HDL	G	0.2593	0.049	2.54E-23	0.0327	0.002136	0.033149	1.26E-05
<i>LIPC</i>	15	rs261338	HDL	A	0.2205	0.0517	3.35E-21	0.0225	0.055035	0.023354	0.006117
<i>LPL</i>	8	chr8:19868772	HDL	G	0.3227	0.0399	3.98E-19	0.0377	3.42E-10	0.037656	9.08E-17
<i>LPL</i>	8	chr8:19959564	HDL	G	0.3155	0.0302	1.88E-11	0.0271	4.02E-06	0.027087	2.11E-09
<i>APOE</i>	19	rs769449	LDL	A	0.2098	0.2097	9.61E-32	0.1671	7.52E-18	0.1739	6.03E-22
<i>APOE</i>	19	chr19:50103919	LDL	A	0.1762	-0.484	9.00E-107	-0.4606	2.02E-79	-0.4603	1.17E-93
<i>PVRL2</i>	19	rs7254892	LDL	A	0.1303	-0.4105	3.84E-34	-0.3762	1.27E-19	-0.39478	3.45E-31
<i>PVRL2</i>	19	rs6857	LDL	A	0.2414	0.2042	1.37E-35	0.1912	1.25E-26	0.196799	6.18E-33
<i>TOMM40</i>	19	rs2075650	LDL	G	0.2243	0.2018	3.40E-33	0.1917	3.04E-26	0.205764	2.38E-34
<i>TOMM40</i>	19	chr19:50092587	LDL	A	0.2384	-0.3502	2.98E-24	-0.3207	3.36E-15	-0.35984	3.56E-25
<i>LDLR</i>	19	rs6511720	LDL	A	0.202	-0.1922	3.59E-27	-0.1868	2.58E-19	-0.18776	1.08E-25
<i>LDLR</i>	19	rs2738447	LDL	A	0.4291	-0.0608	7.68E-07	-0.0525	0.000205	-0.05341	1.48E-05
<i>CELSR2</i>	1	chr1:109618715	LDL	G	0.2823	-0.1502	7.73E-26	-0.1264	5.73E-08	-0.12102	3.09E-09
<i>CELSR2</i>	1	chr1:109618768	LDL	G	0.3629	-0.1127	5.66E-19	-0.0279	0.17683	-0.0364	0.044382
<i>APOB</i>	2	rs41288783	LDL	A	0.2504	0.9996	1.36E-06	0.9255	1.46E-05	0.993053	4.29E-05
<i>APOB</i>	2	rs1367117	LDL	A	0.369	0.1163	1.15E-20	0.0819	2.95E-08	0.090326	3.44E-12
<i>APOB</i>	2	chr2:21124828	LDL	A	0.2278	-0.1629	7.52E-23	-0.1286	1.93E-11	-0.13141	2.64E-14
<i>ZNF259</i>	11	rs6589566	TG	G	0.2199	0.1362	7.91E-27	0.1458	8.76E-36	0.144344	1.30E-29
<i>ZNF259</i>	11	chr11:116160810	TG	A	0.2184	0.1387	7.02E-27	0.1485	3.61E-35	0.146956	9.83E-30
<i>APOA5</i>	11	chr11:116167541	TG	A	0.2187	0.1359	1.74E-26	0.1449	1.71E-34	0.144304	3.62E-29
<i>APOA5</i>	11	chr11:116168917	TG	G	0.211	0.1444	4.18E-27	0.1539	7.29E-34	0.153089	6.94E-30
<i>BUD13</i>	11	chr11:116137710	TG	A	0.2205	0.1311	1.19E-25	0.1408	3.76E-32	0.139785	1.03E-28

<i>BUD13</i>	11	rs10790162	TG	A	0.2194	0.1356	1.23E-26	0.1452	2.02E-35	0.14429	1.43E-29
<i>LOC283143</i>	11	chr11:116031077	TG	A	0.2293	0.09	2.32E-14	0.1855	1.69E-16	0.154655	6.24E-23
<i>LOC283143</i>	11	chr11:116038377	TG	C	0.3288	0.0775	2.34E-06	-0.442	1.67E-27	0.490813	2.87E-85
<i>LOC283143</i>	11	chr11:116063637	TG	A	0.3275	0.0442	2.25E-09	0.0611	1.41E-07	0.036298	6.22E-05
<i>LOC283143</i>	11	chr11:116083427	TG	G	0.3533	0.0507	6.44E-06	-0.1791	1.01E-17	0.317944	3.31E-81
<i>LOC283143</i>	11	rs7350481	TG	A	0.212	0.1193	2.62E-19	-0.1867	1.92E-05	-0.06491	0.026156
<i>LOC283143</i>	11	chr11:116099040	TG	A	0.1983	0.1556	1.36E-13	0.2278	2.07E-11	0.427106	4.90E-47
<i>LOC283143</i>	11	chr11:116109242	TG	G	0.2243	0.1175	1.46E-22	-0.1986	8.31E-07	0.026689	0.369999
<i>LOC283143</i>	11	chr11:116112578	TG	A	0.2171	0.1298	1.16E-24	0.319	1.03E-18	0.183769	1.55E-12
<i>LOC283143</i>	11	chr11:116115504	TG	G	0.3046	-0.0494	1.13E-08	-0.0685	9.05E-11	-0.05613	1.92E-08
<i>LOC283143</i>	11	chr11:116116667	TG	A	0.2301	0.1178	1.19E-16	0.3036	1.15E-13	0.221831	1.15E-22
<i>LOC283143</i>	11	chr11:116117338	TG	G	0.2297	0.1125	2.27E-21	0.1353	7.93E-06	0.146775	3.02E-07
<i>LPL</i>	8	rs10096633	TG	A	0.2527	-0.0939	1.39E-23	-0.0606	6.76E-10	-0.06653	7.81E-11
<i>LPL</i>	8	chr8:19914551	TG	A	0.3402	-0.074	1.26E-24	-0.0527	2.80E-13	-0.05441	3.59E-12
<i>APOE</i>	19	rs769449	TC	A	0.2102	0.2101	2.87E-26	0.1397	3.55E-09	0.157471	2.54E-14
<i>APOE</i>	19	chr19:50103919	TC	A	0.1763	-0.3898	2.38E-56	-0.4014	4.10E-44	-0.38812	9.75E-53
<i>APOE</i>	19	rs439401	TC	A	0.3931	-0.0605	9.87E-06	-0.0827	1.03E-06	-0.06557	4.79E-06
<i>LDLR</i>	19	rs6511720	TC	A	0.2019	-0.2254	7.35E-30	-0.2194	2.68E-21	-0.2206	3.44E-28
<i>LDLR</i>	19	rs2738447	TC	A	0.4287	-0.0664	1.31E-06	-0.0563	0.000179	-0.05779	2.57E-05
<i>PVRL2</i>	19	rs7254892	TC	A	0.131	-0.3109	1.05E-16	-0.2778	4.58E-10	-0.29576	4.04E-15
<i>PVRL2</i>	19	rs6857	TC	A	0.2435	0.1949	1.05E-26	0.1852	3.83E-20	0.189344	3.29E-25
<i>APOB</i>	2	rs41288783	TC	A	0.2021	1.0301	8.06E-07	0.9495	2.16E-05	1.02417	1.40E-05
<i>APOB</i>	2	rs1367117	TC	A	0.3691	0.1195	6.62E-18	0.0842	6.50E-08	0.092005	1.54E-10
<i>APOB</i>	2	rs541041	TC	G	0.2491	-0.1635	7.42E-22	-0.133	9.29E-13	-0.13569	1.25E-14

Table 5-5. Independent effect SNPs identified by HAPRAP in four lipids traits.

Locus	Chr	independent SNP	traits	Allele	Freq	Beta	P	HAPRAP Beta	HAPRAP P	GCTA Beta	GCTA P
<i>F10</i>	13	rs563964	FVII	A	0.4867	4.9929	1.42E-34	4.2915	2.07E-14	4.30832	3.01E-24
<i>F10</i>	13	rs547138	FVII	A	0.5802	3.9481	1.09E-20	2.8342	9.25E-07	2.82781	1.24E-10
<i>ABO</i>	9	chr9:135138971	FVIII	A	0.2984	16.206	1.49E-36	19.6925	1.28E-42	18.6836	7.26E-46
<i>ABO</i>	9	rs507666	FVIII	A	0.3517	14.3829	7.07E-74	16.0027	1.26E-70	15.5104	3.64E-82
<i>OBP2B</i>	9	rs12554449	FVIII	A	0.4635	5.6209	5.41E-06	11.8967	9.28E-16	9.82115	1.27E-14
<i>OBP2B</i>	9	chr9:135112913	FVIII	G	0.2947	15.9482	2.54E-35	9.5292	9.64E-10	13.2242	1.11E-22
<i>OBP2B</i>	9	chr9:135118367	FVIII	A	0.4615	-7.5723	1.17E-31	-14.4936	2.72E-51	-11.0948	2.05E-52
<i>OBP2B</i>	9	chr9:135118593	FVIII	G	0.3745	-5.3952	4.94E-13	-10.2517	1.06E-26	-9.19467	5.49E-29
<i>ADAMTS13</i>	9	chr9:135282263	FVIII	A	0.3108	13.2123	6.50E-08	15.3184	2.07E-09	14.2103	7.82E-09
<i>ADAMTS13</i>	9	chr9:135294712	FVIII	G	0.2858	10.6838	6.04E-12	12.5936	7.69E-13	11.6782	7.58E-14
<i>ADAMTS13</i>	9	rs4962153	FVIII	A	0.3336	9.4932	3.28E-24	10.4451	5.90E-26	10.0274	1.48E-26
<i>CACFD1</i>	9	chr9:135319775	FVIII	A	0.3352	9.4455	6.10E-24	9.5931	1.84E-21	9.5006	3.98E-24
<i>CACFD1</i>	9	chr9:135323177	FVIII	A	0.2165	15.3003	5.37E-06	16.7279	1.97E-05	15.7311	3.12E-06
<i>GBGT1</i>	9	chr9:135032145	FVIII	A	0.3639	5.0703	3.71E-11	6.2262	2.47E-12	5.94377	1.50E-14
<i>GBGT1</i>	9	chr9:135055347	FVIII	A	0.2985	10.7287	7.37E-21	12.117	2.72E-21	11.7318	6.86E-24
<i>ABO</i>	9	rs8176725	vWF	A	0.2909	0.127	2.03E-47	0.1334	3.10E-49	0.131433	1.89E-47
<i>ABO</i>	9	chr9:135123327	vWF	A	0.3427	-0.0661	1.98E-26	-0.0681	6.32E-19	-0.07385	2.81E-28
<i>ABO</i>	9	chr9:135126784	vWF	A	0.4653	-0.0982	4.34E-85	-0.0812	1.24E-26	-0.08735	5.44E-46
<i>ABO</i>	9	chr9:135142543	vWF	A	0.3462	-0.0683	7.66E-29	-0.0879	8.96E-15	-0.07837	2.31E-20
<i>ABO</i>	9	chr9:135146954	vWF	A	0.4419	0.0731	7.12E-49	0.123	1.94E-55	0.10824	2.35E-70
<i>ABO</i>	9	chr9:135171669	vWF	G	0.3879	-0.0555	5.55E-24	-0.0402	2.26E-11	-0.04406	6.97E-15
<i>OBP2B</i>	9	chr9:135106483	vWF	A	0.2786	-0.0439	3.59E-07	-0.0612	4.49E-06	-0.06161	2.83E-10
<i>OBP2B</i>	9	chr9:135111124	vWF	A	0.31	0.0755	5.82E-25	-0.091	6.53E-09	-0.02766	0.00478
<i>OBP2B</i>	9	chr9:135118367	vWF	A	0.4551	-0.0945	1.03E-77	-0.1723	1.88E-123	-0.1428	7.27E-129
<i>OBP2B</i>	9	chr9:135118379	vWF	G	0.2806	0.1838	2.85E-74	0.1443	2.19E-15	0.141907	6.01E-30

<i>OBP2B</i>	9	chr9:135118558	vWF	C	0.3619	-0.076	1.58E-38	-0.17	1.26E-100	-0.1399	2.07E-94
<i>SURF4</i>	9	chr9:135225708	vWF	A	0.285	-0.0589	2.71E-13	-0.0369	7.95E-06	-0.03924	1.72E-06
<i>SURF4</i>	9	chr9:135227493	vWF	A	0.3191	0.1338	2.41E-77	0.1291	4.25E-82	0.128935	4.86E-68
<i>C9orf96</i>	9	chr9:135244970	vWF	A	0.3044	-0.0601	9.39E-15	-0.0656	9.88E-16	-0.05218	2.94E-10
<i>C9orf96</i>	9	chr9:135254869	vWF	A	0.2796	-0.0545	2.33E-09	-0.0627	3.28E-11	-0.05006	1.25E-07
<i>C9orf96</i>	9	chr9:135256970	vWF	A	0.3159	0.1284	1.42E-68	0.103	1.82E-21	0.087128	9.29E-19
<i>C9orf96</i>	9	chr9:135257905	vWF	A	0.403	0.04	8.99E-14	-0.0437	6.80E-10	-0.01397	0.024119
<i>C9orf96</i>	9	chr9:135258289	vWF	C	0.3085	-0.0364	7.95E-07	-0.0413	1.01E-07	-0.03377	1.56E-05
<i>C9orf96</i>	9	chr9:135260359	vWF	C	0.2954	0.138	1.18E-55	0.0607	3.04E-07	0.066674	2.04E-09
<i>ADAMTS13</i>	9	chr9:135282263	vWF	A	0.2832	0.1803	5.03E-21	0.2063	5.97E-28	0.192827	3.43E-23
<i>ADAMTS13</i>	9	chr9:135294712	vWF	G	0.2728	0.106	1.91E-17	0.131	2.27E-25	0.119081	2.67E-21
<i>ADAMTS13</i>	9	rs4962153	vWF	A	0.3155	0.1248	8.98E-66	0.1357	5.01E-77	0.130568	1.82E-70
<i>CACFD1</i>	9	rs739468	vWF	A	0.3141	0.1241	2.83E-65	0.1255	1.99E-70	0.124649	4.92E-65
<i>CACFD1</i>	9	chr9:135323177	vWF	A	0.2176	0.1434	2.62E-07	0.1619	8.25E-09	0.150061	8.16E-08
<i>MED22</i>	9	rs621907	vWF	A	0.276	-0.0497	1.48E-07	-0.044	4.72E-06	-0.04566	1.60E-06
<i>MED22</i>	9	chr9:135201989	vWF	A	0.2708	0.1688	1.81E-32	0.1653	1.20E-30	0.166274	5.45E-31
<i>SLC2A6</i>	9	chr9:135326625	vWF	A	0.287	0.1811	3.99E-21	0.1763	5.26E-21	0.177274	4.93E-19
<i>SLC2A6</i>	9	chr9:135328419	vWF	A	0.2654	-0.0623	1.63E-10	-0.0596	1.95E-08	-0.06953	2.09E-12
<i>SLC2A6</i>	9	chr9:135329780	vWF	G	0.3176	-0.0421	3.94E-09	-0.0649	3.86E-05	-0.03313	0.008265
<i>SLC2A6</i>	9	chr9:135330021	vWF	A	0.3823	0.0437	8.94E-15	0.096	8.40E-39	0.077087	1.94E-31
<i>SLC2A6</i>	9	chr9:135333468	vWF	C	0.3361	-0.0512	7.70E-15	-0.06	8.89E-06	-0.07193	9.20E-11
<i>SLC2A6</i>	9	chr9:135337653	vWF	G	0.2576	-0.0502	2.19E-06	-0.0479	1.32E-05	-0.05566	2.23E-07
<i>SLC2A6</i>	9	chr9:135338906	vWF	C	0.341	-0.0375	3.18E-09	-0.0276	1.95E-05	-0.03408	1.05E-07
<i>GBGT1</i>	9	chr9:135032145	vWF	A	0.3536	0.0588	1.90E-22	0.071228	1.26E-31	0.067896	4.75E-29
<i>GBGT1</i>	9	chr9:135055347	vWF	A	0.2828	0.1144	1.97E-36	0.130282	7.80E-45	0.125969	1.35E-41

Table 5-6. Independent effect SNPs identified by HAPRAP in three coagulation traits.

5.3.5. Fine mapping for other CVD related traits

I run the fine mapping analysis using UCLEB meta-analysis results of other CVD related traits. For il6 and surate, I did not find any additional signals except the top hits rs7518199 (associated with il6) and rs9998811 (associated with surate).

As shown in Table 5-7, totally 10 SNPs in 4 loci have multiple association SNPs for bilirubin and crp: 3 SNPs in UGT region associated with bilirubin and 7 SNPs in 3 loci associated with crp.

Locus	Chr	independent SNP	traits	Allele	Freq	Beta	P	HAPRAP Beta	HAPRAP P	GCTA Beta	GCTA P
<i>UGT</i>	2	rs28899170	bilirubin	A	0.4362	2.4878	9.74E-157	1.2848	1.29E-13	1.29489	2.54E-12
<i>UGT</i>	2	rs17862875	bilirubin	A	0.4291	2.5304	5.68E-159	1.3552	9.75E-15	1.37247	2.24E-13
<i>UGT</i>	2	rs11568318	bilirubin	A	0.3609	-1.5878	1.31E-17	-0.7261	3.71E-05	-1.22312	1.21E-10
<i>CRP</i>	1	rs1205	crp	A	0.3857	-0.1662	1.60E-26	-0.0781	2.44E-05	-0.08849	1.30E-06
<i>CRP</i>	1	rs3091244	crp	A	0.4039	0.1865	1.33E-34	0.1412	3.77E-15	0.140156	3.72E-15
<i>CRP</i>	1	rs11579148	crp	A	0.1826	-0.3538	2.16E-08	-0.247	4.73E-05	-0.31971	6.32E-07
<i>HNF1A</i>	12	chr12:119901371	crp	G	0.3712	-0.1649	1.20E-25	-0.1834	3.87E-30	-0.17498	2.29E-28
<i>HNF1A</i>	12	rs2245407	crp	A	0.197	-0.1345	4.47E-06	-0.1959	1.98E-10	-0.17043	7.74E-09
<i>C12orf43</i>	12	chr12:119926624	crp	A	0.187	-0.1428	6.71E-06	-0.1957	6.41E-09	-0.17379	5.01E-08
<i>C12orf43</i>	12	chr12:119927499	crp	G	0.3574	-0.1581	6.83E-23	-0.1729	5.46E-27	-0.16637	5.30E-25

Table 5-7. Independent effect SNPs identified by HAPRAP in *crp* and *bilirubin*.

5.3.6. Fine mapping for Atopic Dermatitis

For the fine mapping analysis of Atopic Dermatitis, I applied HAPRAP to meta-analysis results. Four genomic regions with more than one independent SNPs were determined: six SNPs were identified in chromosome 1 (region from 150250000 to 154500000); two independent SNPs in chromosome 4 (region from 122993592 to 123493592), chromosome 5 (region from 131741085 to 132150000) and chromosome 10 (region from 6020000 to 6288853) (Table 5-21). Other twelve regions showed no evidence for secondary signals.

Chr	SNP	beta	SE	pval	N	HAPRAP beta	HAPRAP SE	HAPRAP pval
Chr1	1q21.2_SNP1	-0.0771	0.012585032	9.20E-10	102760	-0.06658198	0.0125	1.00301E-07
Chr1	1q21.3_SNP1	0.2755	0.030093897	5.75E-20	96795	0.162593083	0.033	8.36123E-07
Chr1	1q21.3_SNP2	-0.1139	0.015825028	6.31E-13	102760	-0.07749054	0.0161	1.48835E-06
Chr1	1q21.3_SNP3	0.4764	0.042608896	5.55E-29	93326	0.469246255	0.0428	5.94072E-28
Chr1	1q21.3_SNP4	0.2961	0.031383212	4.12E-21	96795	0.169504985	0.035	1.28091E-06
Chr1	1q21.3_SNP5	0.0783	0.012288311	1.93E-10	102760	0.073674469	0.0122	1.55707E-09
Chr4	4q27_SNP1	0.0758	0.012616106	1.94E-09	102761	0.054018999	0.0135079	6.36371E-05
Chr4	4q27_SNP2	-0.0897	0.015254395	4.20E-09	102762	-0.0673655	0.0163327	3.71707E-05
Chr5	5q31.1_SNP1	0.1271	0.015100459	4.03E-17	102761	0.10440521	0.018192895	9.56E-09
Chr5	5q31.1_SNP2	0.0986	0.014370354	7.14E-12	101294	0.072118769	0.016450248	1.17E-05
Chr10	10p15.1_SNP1	0.0736	0.012177861	1.54E-09	103065	0.058452566	0.014549342	5.89E-05
Chr10	10p15.1_SNP2	0.1778	0.03176829	2.23E-08	97099	0.146938916	0.034499725	2.05E-05

Table 5-21. Genomic regions with multiple SNPs associated with Atopic Dermatitis using fixed effect model meta-analysis results. I used chromosome location instead of the SNP ID here since the results are unpublished data.

5.4. Discussion

5.4.1. Fine mapping of *ABCG5-8* region

The analysis I conducted for *ABCG5* and *8* region associated with GBD replicated the conclusion in the previous meta-analysis (Rodriguez et al., 2014). Rs4953023, which is the top hit of the GBD meta-analysis, is in perfect LD with a previously associated SNP rs11887534 (Buch et al., 2007). Rs4299376 is only 1424 base pair apart from the top hit. These two SNPs are in weak LD with r^2 equal to 0.047 (1000 Genome CEU population).

5.4.2. Fine mapping of *NOS1AP* region

For *NOS1AP* associated with QTc interval, the first independent SNP rs10429888 is in strong LD with a previously associated SNP rs12143842 ($r^2=0.646$ in 1000 Genome CEU), this SNP was indicated as the most significant SNP in two previous reports (Newton-Cheh et al., 2009; Pfeufer et al., 2009). Meanwhile, the second independent SNP rs10918859 is in moderate LD with two previously reported SNPs rs4657178 ($r^2=0.467$ in 1000 Genome CEU) and rs12029454 ($r^2=0.168$ in 1000 Genome CEU). Moreover, conditional on rs12143842, both were reported to be secondary signals in separate studies: rs4657178 in Pfeufer's study (Pfeufer et al., 2009) and rs12029454 in Newton-Cheh's study (Newton-Cheh et al., 2009). In addition, there is evidence of correlations between the third independent SNP, rs16057031, and two different previously reported SNPs rs10494366 ($r^2=0.168$ in 1000 Genome CEU) and rs288058 ($r^2=0.152$ in 1000 Genome CEU) (Arking et al., 2006; Marroni et

al., 2009). SNP rs10494366 was the first well-replicated *NOS1AP* SNP reported to be associated with QT interval.

5.4.3. Fine mapping of *SCN5A-10A* region

For *SCN5A-10A* region associated with PR interval, the top associated SNP rs6801957 was identified as significant SNP by a previous GWAS study (Butler et al., 2012). Other two SNPs are not included in 1000 Genome project, but one of them, rs73056438 was genotyped by the ALSPAC in 8365 individuals. So I used ALSPAC individual-level genotype data to calculate the LD between rs73056438 and other 6 SNPs reported to be associated with PR interval in *SCN5A-10A* region (the 6 SNPs were listed in Table 8-2). None of the previous reported SNPs were in LD with this SNP which suggest that it may represent a novel independent associated SNP.

5.4.4. Fine mapping of candidate regions associated with Atopic Dermatitis

As listed in the introduction section, previous GWAS have identified 28 SNPs associated with Atopic Dermatitis (Table 8-21). The region of chromosome 1 where HAPRAP identifies 6 independent signals is a locus known to be associated with atopic dermatitis. Several rare loss of function FLG mutations have been shown to be strongly associated with atopic dermatitis (Palmer et al., 2006) (Rodríguez et al., 2009) and previous GWAS signals in this region have been shown to be tagging these Loss-of-Function variants (Paternoster et al., 2012). The variants identified by HAPRAP may therefore tag the haplotypes that these rare FLG mutations fall on, or could represent additional independent signals. The EAGLE

consortium is carrying out further analysis in this region to determine if there are any novel independent associations. In chromosome 4, there was only one SNP, rs17389664, was found to be associated with atopic dermatitis (Ellinghaus et al., 2013). In this study, I found that 4q27_SNP1 is in modest LD to rs17389664 (ALSAPC $r^2=0.55$). There is almost no LD between the other SNP 4q27_SNP2 and rs17389664 (ALSAPC $r^2=0.079$), which suggest a secondary signal in this region. On the other hand, there are 2 known independent signals in chromosome 5: rs2897442 and rs848 (Paternoster et al., 2012). In this study, I found two proxy SNPs which replicated these two signals: 5q31.1_SNP1 is in modest LD with rs848 (ALSAPC $r^2=0.281$), 5q31.1_SNP2 is in modest LD with rs848 (ALSAPC $r^2=0.186$) and 5q31.1_SNP2 is in strong LD with rs2897442 (ALSAPC $r^2=0.912$). Moreover, two independent SNPs in chromosome 10 are novel findings in this meta-analysis. Further replication analysis is needed to test these two signals.

The SNPs consisted with previous findings confirm the accuracy of the meta-analysis results and fine mapping using HAPRAP program. Further analysis is necessary for new findings in chromosome 1, 4 and 10.

5.4.5. Conclusion

In this chapter, I focused on applying the method I developed, HAPRAP, to meta-analysis results of several quantitative traits and diseases outcomes. The fine mapping of ECG traits and GBD outcomes are highly consistent to previous conditional analysis of meta-analysis. The fine mapping of blood lipids traits, coagulations traits and other CVD related traits found some novel independent effect SNPs. The fine mapping of atopic dermatitis outcomes

confirmed the previous evidences and enhanced the understanding of its genetic and molecular mechanisms of atopic dermatitis.

CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTION

There has been an increasing number of national and international collaborations in the field of complex disease genetics. The collaboration of large consortia have resulted in an increasing number of large-scale meta-analyses to be carried out for a range of complex traits such as BMI (Speliotes et al, 2010), lipids traits (Rodriguez et al, 2015) and ECG traits (Gaunt et al, 2012). These meta-analyses have successfully increased the statistical power to detect novel genetic variants associated with complex human traits. Moreover, an increasing number of the world's leading international consortia such as the GIANT consortium, are willing to share their meta-analysis summary statistics, either through making the data publicly available or by sharing the data in a collaborative basis. Table 6-1, I lists a set of consortia which provide meta-analysis summary data publicly. These summary data are unique and informative resources which can lead us to better understand the genetic basis of different diseases/traits.

Trait	Consortium / database	Reference	Sample size
Age-related macular degeneration (AMD)	AMD Gene Consortium	Fritsche et al, 2013, Nat Genet	77100
Rheumatoid Arthritis	BIRAC and YEAR	Stahl et al, 2010, Nat Genet	25708
Coronary artery disease	CARDIOGRAM	Schunkert, et al, 2011, Nat Genet	76995
Atopic Dermatitis	EAGLE	Lavinia, et al, 2015 Nat Genet	40835
Birth Head Circumference	EGG	Taal, et al, 2012, Nat Genet	10768
Asthma	GABRIEL	Moffatt, et al, NEJM, 2010	26475
BMI	GIANT	Yang, et al, 2012, Nature	169811
Extrem_BMI	GIANT	Berndt, et al, 2013, Nat Genet	263407
Extrem_height	GIANT	Berndt, et al, 2013, Nat Genet	263407
Extrem_WHR	GIANT	Berndt, et al, 2013, Nat Genet	263407
Height	GIANT	Wood, et al, 2014, Nat Genet	253288
hip circumference	GIANT	Shungin, et al, 2015, Nature	224459

Obesity	GIANT	Berndt, et al, 2013, Nat Genet	263407
Overweight	GIANT	Berndt, et al, 2013, Nat Genet	263407
waist circumference	GIANT	Shungin, et al, 2015, Nature	224459
waist-to-hip ratio (WHR)	GIANT	Shungin, et al, 2015, Nature	224459
Lipids	GLGC	Willer, et al, 2013, Nat Genet	188577
Crohn's disease.	IBD	Franke, et al, 2010, Nat Genet	21389
Ulcerative colitis	IBD	Anderson, et al, 2011, Nat Genet	25865
Celiac Disease	ImmunoBase	Trynka, et al, 2011, Nat Genet	24242
Juvenile Idiopathic Arthritis	ImmunoBase	Hinks, et al, 2013, Nat Genet	15872
narcolepsy	ImmunoBase	Faraco, et al, 2013, PLoS Genet	12307
Primary Biliary Cirrhosis	ImmunoBase	Liu JZ, et al, 2012, Nat Genet	11375
Psoriasis	ImmunoBase	Tsoi LC, et al, 2012, Nat Genet	33394
	ImmunoBase /	International Multiple Sclerosis Genetics	
Multiple Sclerosis	IMSGC	Consortium, 2013, Nat Genet	38589
Metabolites	MuTHER	Shin, et al, 2014, Nat Genet	7824
Sporadic amyotrophic lateral			
sclerosis	SLAGEN	Fogh, et al, 2014, HMG	13225
Type 1 Diabetes	T1D Base	Copper, Unpublished	13254

Table 6-1. *A list of consortium with public available meta-analysis summary data*

As explained in Chapter 1, utilising publicly available summary statistics has become widespread in the complex trait genetics field. The below are some of the well-known and widely used examples – as mentioned in the previous chapter:

1. LD score regression using meta-analysis summary statistics for heritability analysis (Bulik-Sullivan BK et al, 2015 a) and genetic correlation analysis (Bulik-Sullivan BK et al, 2015 b).
2. Mendelian randomisation approach which can be applied to identify causal relationship using public available data (Burgess et al, 2015).
3. GCTA conditional and joint SNP effect analysis (Yang et al, 2012) using meta-analysis summary statistics to process an approximately conditional analysis

Under such trend of utilising meta-analysis summary statistics, I developed two user-friendly statistical methods, namely SSS-RAP (Zheng et al, 2013) and HAPRAP (Zheng et al, 2015), which are two model selection approaches using meta-analysis summary statistics directly. I then applied these two methods to certain complex traits and compared them to existing state-of-the-art tools. As described in chapter 2 and 4, these two methods make better use of the meta-analysis summary statistics and their accuracy and performance are better than the existing state-of-the-art tools.

6.1. Selecting independent SNPs using Sequential Sentinel SNP Regional Association Plot

In this thesis, I designed and developed the Sequential Sentinel SNP Regional Association Plot (SSS-RAP) program (Zheng et al, 2013) which utilises the concept of pair-wise LD to calculate SNP effects conditional on the most significant signals (defined as ones with the lowest p values). It was our first try to filling the holes present in fine mapping analysis using meta-analysis summary results. The performance of SSS-RAP is comparable to other model selection methods at individual-level. Although the method has been published few years ago, it is still valuable since 1) it only requires summary-level meta-analysis data and summary LD information, which is an advantage over all other model selection methods at current; 2) it is computationally fast, which only need a few seconds to finish an approximately step-wise regression procedure. So it is a quick, convenient and a visually presented method for testing SNP 'independence' .

6.2. Haplotype based model selection using Haplotype regional association analysis Program.

The bulk of this thesis concerns itself with the design and development of Haplotype regional association analysis (HAPRAP). This method incorporates the biological concept of haplotypes and an empirical expectation-maximisation (E-M) algorithm incorporating the relationship between marginal SNP effects and joint SNP effects. The results of HAPRAP were highly consistent with multiple regression results using individual-level data. In meta-analysis level, it is the most accurate approach for identifying SNP independence at current. I summarise below its advantages over other tools, both in biological and statistical point of view:

In a biological point of view, HAPRAP inherits the core concept of using haplotypes to represent LD from SSS-RAP and extend it to multiple loci situation. The core advantage of HAPRAP is that it is the first approach to insert haplotype based analysis concepts to SNP based analysis and utilizes this nested design to identify SNP independence – using meta-analysis summary results. In fact, in individual level, haplotype based association analysis has already proved its advantage against single marker analysis. As demonstrated by previous simulation studies, statistical approaches based on haplotypes can be a more powerful method to characterize the genetic background of complex diseases compare to single maker LD analysis (Akey J et al, 2001). In meta-analysis level, HAPRAP, showed its accuracy advantage against GCTA-COJO, which inherits the pair-wise correlation concept from multiple regression. Moreover, from a statistical point of view, the core formula of

HAPRAP (see Chapter 5.2.1, equation 2) starts from a very simple concept of a single regression between a SNP and a trait, which assumes an additive model, treats the common allele as the “baseline” allele (with respect to trait) and assumes the rare allele is the “effect” allele. HAPRAP then extends this concept to a more complex SNP-haplotype matrix. However, the model is an EM style algorithm which is empirical. Although it is possible that the EM algorithm may not convergence may not be possible, so I cannot exclude the situation that HAPRAP will not converge. However, in hundreds of thousands of simulation and real case examples we performed, I did not find any situation where HAPRAP did not converge. Algorithms are often used effectively where the biological model is well understood, but the statistical model is too complex to generalise to all scenarios. For instance, a recent fine mapping method, probability identification of causal SNPs (PICS), used an empirical constant in its core algorithm to estimate the expected mean of the association signal at a SNP (Farh et al., 2015). I believe, with further development of haplotype-based analyses - especially with advancements on including haplotypes in appropriate statistical models - the importance of HAPRAP algorithm will increase in the future.

6.3. Future directions

In this section, I will describe possible extension of SSS-RAP and HAPRAP programme and potential extensions of the fine mapping approaches.

6.3.1. Further development of SSS-RAP: fine mapping for untyped SNPs

In an association analysis of a specific region, the real causal SNP may not be directly genotyped because of chosen SNP array coverage. Nowadays, imputation is a widely used method that increases the density of SNP data available from an array. But the quality of imputation is highly dependent on the density of the original genotype data and the quality of the reference panel being used (Marchini and Howie, 2010). The causal SNP may still not be detected, even after imputation. Therefore, to test whether there is an untested and “significant” SNP associated with a trait, and to determine how it relates to directly genotyped SNPs that show significance, I am working on an idea that develops a “hidden sentinel SNP shadow LD pattern” approach using the SSS-RAP program. This method will extend our knowledge of the associations between typed/imputed SNPs in a region and phenotypes, to all SNPs in a genomic region. More importantly, there will be no extra genotyping cost, which makes this method more attractive.

6.3.2. Further development of HAPRAP

As I alluded to before, HAPRAP can be a useful tool in four aspects. In this section, I will describe the possible extension of HAPRAP in these aspects:

1) HAPRAP provides a good platform for the functional annotation of casual variants using variant prediction tools such as FATHMM (Shihab et al., 2015). One of the key purposes of fine mapping is to select the variant that is most likely to be the casual SNP for further functional studies. Therefore, a logical follow-up to HAPRAP would be to find biological links

for those SNPs that are reported to be associated with complex traits (Schaub et al., 2012). Ideally, this process should be an automated procedure and be feasibly carried out on multiple SNPs. To identify possible functional consequences of the independent SNPs, functional prediction tools such as FATHMM (Shihab et al., 2013) and RegulomeDB (Boyle et al., 2012) can be used to functionally annotate the SNPs with independent effects (as identified by HAPRAP). This process can be further extend to SNPs with very high LD (eg: $r^2 \geq 0.8$) to the independent SNPs identified using tools such as SNAP (Johnson et al., 2008).

2) HAPRAP provides an automated model selection procedure, which can be used for allelic score estimating. Although HAPRAP is developed for analysing meta-analysis summary data, it can be used using individual-level too. As mentioned in Chapter 7.2, haplotype-based analysis maintain advantages over multiple regression using individual-level data, so HAPRAP can be used as a convenient automated step-wise selection procedure in individual-level as well. As mentioned in Chapter 1, allele scores (also called genetic risk scores, gene scores, or genotype scores) are a convenient way of summarizing a large number of genetic variants associated with a risk factor. It is calculated as the total number of risk factor increasing alleles for an individual (unweighted score), or the sum of weights for each allele corresponding to estimated genetic effect sizes (weighted score) (Burgess S et al, 2012). The most important procedure of allelic score estimating is to select a subset of SNPs which represents most of the total variance explained while keeping with the stringency criteria. Furthermore, HAPRAP is also a suitable model selection approach for the genome-wide allelic score of biological intermediates and mining the phenome (Evans et al.,

2013), which can be used in Mendelian randomization and will lead to “Hypothesis-Free” causality (Evans and Davey Smith, 2015).

3) HAPRAP provides a good SNP selection procedure for two-sample Mendelian randomisation (MR) approach using publicly available data to estimate the causal effect of the risk factor on an outcome (Burgess et al, 2015). In the current two-sample MR framework, if the test SNPs are in LD with each other, their effects are merged together to build a combined effect using the LD correlation matrix (cite). HAPRAP provides an alternative way, which can be used as a SNP selection procedure and only the best fit model are then be used for the MR. Compared to other model selection approaches, HAPRAP requires the same summary information, which will have been corrected for the two-sample MR already. As presented in Chapter 5.4, HAPRAP is the most accurate method when using public available genotypes (such as the 1000 Genome project data) as a reference panel. So combining HAPRAP with MR will lead to a more accurate and robust causal relationship between two or more traits.

4) HAPRAP’s unique function of identifying SNP independence for rare variants. As explained in Chapter 1, rare variants tends to have larger effects than common variants (Bodmer and Bonilla, 2008). One limitation of rare variant analysis is the lack of a rare variant catalog with reference genotypes. For common variants, publicly available genotype resources such as the HapMap (www.hapmap.org/) and 1000 Genome Project (<http://www.1000genomes.org/>) databases exist. These reference genotypes help identify ‘tag SNPs’ within a group of SNPs in high LD. I believe that, with further additions to the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010) with other publicly

available whole genome sequencing projects, the reference genotypes for rare variants will be better established in the future (Asimit and Zeggini, 2010). For example, the UK10K project (<http://www.uk10k.org/>) is another publicly available whole genome sequencing dataset that can be used as an LD reference. UK10K has a larger sample size than the 1000 Genome datasets, and may have better LD structure (Timpson et al., 2014; Taylor et al., 2015). The novel genotype reference panel for rare variants will provide a good foundation for further statistical and bioinformatics analysis, especially for fine mapping of rare variants.

6.4. Conclusion

For many complex human traits and diseases (e.g. cardiovascular disease, atopic dermatitis), GWAS studies and meta-analysis of multiple GWAS studies have successfully identified multiple loci associated with common traits. Using the two methods I developed, we can carry out fine mapping in genomic regions with possible secondary signals using meta-analysis summary results. As explained in this chapter, future development of this work will focus on three aspects: i) Genetic refinement of association signals, fine mapping of rare variants will be possible in the near future when we have a fuller understanding of haplotype pattern pertaining to rare variants, ii) Phenotypic refinement of association signals, improved phenotype definition by increasing the specificity of phenotype (e.g. detailed metabolic phenotypes instead of lipid traits) which will increase the power of detecting association signals; and iii) Following up association signals with functional analysis, since correlated loci are not necessarily a biological agent for disease. Functional data, such as expression information and functional annotation, can certainly support and

inform the efforts of finding casual signals. New methodologies and approaches in these three aspects will improve our understanding of common complex human diseases.

CHAPTER 7. REFERENCE

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073.

Acharya, S.S., Coughlin, A., Dimichele, D.M., and North American Rare Bleeding Disorder Study Group (2004). Rare Bleeding Disorder Registry: deficiencies of factors II, V, VII, X, XIII, fibrinogen and dysfibrinogenemias. *J. Thromb. Haemost. JTH* 2, 248–256.

Agerholm-Larsen, B., Tybjaerg-Hansen, A., Schnohr, P., Steffensen, R., and Nordestgaard, B.G. (2000). Common cholesteryl ester transfer protein mutations, decreased HDL cholesterol, and possible decreased risk of ischemic heart disease: The Copenhagen City Heart Study. *Circulation* 102, 2197–2203.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions Autom. Control* 19, 716–723.

Akey J, Jin L, Xiong M: Haplotypes vs single marker linkage disequilibrium tests: what do we gain?

Eur J Hum Genet 2001, 9(4):291-300.

Aledort, L., Ljung, R., Mann, K., and Pipe, S. (2014). Factor VIII therapy for hemophilia A: current and future issues. *Expert Rev. Hematol.* 7, 373–385.

Alharbi, K.K., Aldahmesh, M.A., Spanakis, E., Haddad, L., Whittall, R.A., Chen, X., Rassoulilian, H., Smith, M.J., Sillibourne, J., Ball, N.J., et al. (2005). Mutation scanning by meltMADGE: validations using BRCA1 and LDLR, and demonstration of the potential to identify severe, moderate, silent, rare, and paucimorphic mutations in the general population. *Genome Res.* 15, 967–977.

Alharbi, K.K., Aldahmesh, M.A., Gaunt, T.R., Rassoulilian, H., Ai Guthrie, P., Rodriguez, S., Boustred, C.R., Spanakis, E., and Day, I.N.M. (2010). MeltMADGE for mutation scanning of specific genes in population studies. *Nat. Protoc.* 5, 1800–1812.

Antonarakis, S.E., Youssoufian, H., and Kazazian, H.H., Jr (1987). Molecular genetics of hemophilia A in man (factor VIII deficiency). *Mol. Biol. Med.* 4, 81–94.

Arking, D.E., Pfeufer, A., Post, W., Kao, W.H.L., Newton-Cheh, C., Ikeda, M., West, K., Kashuk, C., Akyol, M., Perz, S., et al. (2006). A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat. Genet.* 38, 644–651.

Asimit, J., and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* 44, 293–308.

Asselbergs, F.W., Guo, Y., van Iperen, E.P.A., Sivapalaratnam, S., Tragante, V., Lanktree, M.B., Lange, L.A., Almoguera, B., Appelman, Y.E., Barnard, J., et al. (2012). Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am. J. Hum. Genet.* 91, 823–838.

- Bachmann, J.M., Willis, B.L., Ayers, C.R., Khera, A., and Berry, J.D. (2012). Association between family history and coronary heart disease death across long-term follow-up in men: the Cooper Center Longitudinal Study. *Circulation* 125, 3092–3098.
- Bainton, D., Miller, N.E., Bolton, C.H., Yarnell, J.W., Sweetnam, P.M., Baker, I.A., Lewis, B., and Elwood, P.C. (1992). Plasma triglyceride and high density lipoprotein cholesterol as predictors of ischaemic heart disease in British men. The Caerphilly and Speedwell Collaborative Heart Disease Studies. *Br. Heart J.* 68, 60–66.
- Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nat. Genet.* 40, 340–345.
- Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinforma. Oxf. Engl.* 21, 263–265.
- Becker, M.A., and Jolly, M. (2006). Hyperuricemia and associated diseases. *Rheum. Dis. Clin. North Am.* 32, 275–293, v–vi.
- Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biom. J. Biom. Z.* 52, 708–721.
- Bennet, A.M., Di Angelantonio, E., Ye, Z., Wensley, F., Dahlin, A., Ahlbom, A., Keavney, B., Collins, R., Wiman, B., de Faire, U., et al. (2007). Association of apolipoprotein E genotypes with lipid levels and coronary risk. *JAMA J. Am. Med. Assoc.* 298, 1300–1311.
- Bieber, T. (2008). Atopic dermatitis. *N. Engl. J. Med.* 358, 1483–1494.
- Biesecker, L.G., Burke, W., Kohane, I., Plon, S.E., and Zimmern, R. (2012). Next-generation sequencing in the clinic: are we ready? *Nat. Rev. Genet.* 13, 818–824.
- Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* 72, 291–336.
- Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* 40, 695–701.
- Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilit. *Pubblicazioni R Ist. Super. Sci. Econ. E Commer. Firenze* 8, 3–62.
- Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.
- Boyd, A., Golding, J., Macleod, J., Lawlor, D.A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., and Davey Smith, G. (2013). Cohort Profile: the “children of the 90s”--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* 42, 111–127.
- Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB.

Genome Res. 22, 1790–1797.

Breslow, J.L., Zannis, V.I., SanGiacomo, T.R., Third, J.L., Tracy, T., and Glueck, C.J. (1982). Studies of familial type III hyperlipoproteinemia using as a genetic marker the apoE phenotype E2/2. *J. Lipid Res.* 23, 1224–1235.

Brown, M.S., and Goldstein, J.L. (1986). A receptor-mediated pathway for cholesterol homeostasis. *Science* 232, 34–47.

Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097.

Browning, S.R., and Browning, B.L. (2011). Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 12, 703–714.

Buch, S., Schafmayer, C., Völzke, H., Becker, C., Franke, A., von Eller-Eberstein, H., Kluck, C., Bässmann, I., Brosch, M., Lammert, F., et al. (2007). A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. *Nat. Genet.* 39, 995–999.

Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG; EPIC- InterAct Consortium. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol.* 2015 Mar 15.

Butler, A.M., Yin, X., Evans, D.S., Nalls, M.A., Smith, E.N., Tanaka, T., Li, G., Buxbaum, S.G., Whitsel, E.A., Alonso, A., et al. (2012). Novel Loci Associated With PR Interval in a Genome-Wide Association Study of 10 African American Cohorts. *Circ. Cardiovasc. Genet.* 5, 639–646.

Buyske, S., Wu, Y., Carty, C.L., Cheng, I., Assimes, T.L., Dumitrescu, L., Hindorff, L.A., Mitchell, S., Ambite, J.L., Boerwinkle, E., et al. (2012). Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study. *PloS One* 7, e35651.

Cambon-Thomsen, A. (2003). Assessing the impact of biobanks. *Nat. Genet.* 34, 25–26.

Cardon, L.R., and Bell, J.I. (2001). Association study designs for complex diseases. *Nat. Rev. Genet.* 2, 91–99.

Castelli, W.P. (1988). Cholesterol and lipids in the risk of coronary artery disease--the Framingham Heart Study. *Can. J. Cardiol.* 4 Suppl A, 5A–10A.

Chang, K.-C., Barth, A.S., Sasano, T., Kizana, E., Kashiwakura, Y., Zhang, Y., Foster, D.B., and Marbán, E. (2008). CAPON modulates cardiac repolarization via neuronal nitric oxide synthase signaling in the heart. *Proc. Natl. Acad. Sci. U. S. A.* 105, 4477–4482.

- Chattopadhyay, S., and Arora, R. (2014). The ironies of human mind: a case of rett syndrome. *Ethiop. J. Health Sci.* *24*, 171–174.
- Chen, L., Tovar-Corona, J.M., and Urrutia, A.O. (2012). Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *Int. J. Evol. Biol.* *2012*, 596274.
- Civeira, F., Pocoví, M., Cenarro, A., Casao, E., Vilella, E., Joven, J., González, J., Garcia-Otín, A.L., and Ordovás, J.M. (1996). Apo E variants in patients with type III hyperlipoproteinemia. *Atherosclerosis* *127*, 273–282.
- Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., and Lander, E.S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* *104*, 19428–19433.
- Clarke, R., Emberson, J.R., Parish, S., Palmer, A., Shipley, M., Linksted, P., Sherliker, P., Clark, S., Armitage, J., Fletcher, A., et al. (2007). Cholesterol fractions and apolipoproteins as risk factors for heart disease mortality in older men. *Arch. Intern. Med.* *167*, 1373–1378.
- Clementi, M., Di Gianantonio, E., Fabris, L., Forabosco, P., Strazzabosco, M., Tenconi, R., and Okolicsanyi, L. (2007). Inheritance of hyperbilirubinemia: evidence for a major autosomal recessive gene. *Dig. Liver Dis. Off. J. Ital. Soc. Gastroenterol. Ital. Assoc. Study Liver* *39*, 351–355.
- Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* *305*, 869–872.
- Cohen, J.C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G.L., Grundy, S.M., and Hobbs, H.H. (2006). Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl. Acad. Sci. U. S. A.* *103*, 1810–1815.
- Colhoun, H.M., McKeigue, P.M., and Davey Smith, G. (2003). Problems of reporting genetic associations with complex outcomes. *Lancet* *361*, 865–872.
- Collins, F.S., and Mansoura, M.K. (2001). The Human Genome Project. Revealing the shared inheritance of all humankind. *Cancer* *91*, 221–225.
- Collins, F.S., Brooks, L.D., and Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* *8*, 1229–1231.
- Conlan, M.G., Folsom, A.R., Finch, A., Davis, C.E., Sorlie, P., Marcucci, G., and Wu, K.K. (1993). Associations of factor VIII and von Willebrand factor with age, race, sex, and risk factors for atherosclerosis. The Atherosclerosis Risk in Communities (ARIC) Study. *Thromb. Haemost.* *70*, 380–385.
- Cooper, J.K. (1986). Electrocardiography 100 years ago. Origins, pioneers, and contributors. *N. Engl. J. Med.* *315*, 461–464.
- Cooper, D.N., Millar, D.S., Wacey, A., Banner, D.W., and Tuddenham, E.G. (1997). Inherited factor VII deficiency: molecular genetics and pathophysiology. *Thromb. Haemost.* *78*, 151–160.

Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L., and Pericak-Vance, M.A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261, 921–923.

Crick, F. (1970). Central dogma of molecular biology. *Nature* 227, 561–563.

Davey Smith G. and Ebrahim S. (2003), Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 32: 1-22

Day, I.N.M. (2010). dbSNP in the detail and copy number complexities. *Hum. Mutat.* 31, 2–4.

Deary, I.J., Whiteman, M.C., Pattie, A., Starr, J.M., Hayward, C., Wright, A.F., Carothers, A., and Whalley, L.J. (2002). Cognitive change and the APOE epsilon 4 allele. *Nature* 418, 932.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181.

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* 10, 5–6.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–38.

Djoussé, L., Levy, D., Cupples, L.A., Evans, J.C., D'Agostino, R.B., and Ellison, R.C. (2001). Total serum bilirubin and risk of cardiovascular disease in the Framingham offspring study. *Am. J. Cardiol.* 87, 1196–1200; A4, 7.

Duret, L. (2009). Mutation patterns in the human genome: more variable than expected. *PLoS Biol.* 7, e1000028.

Eddy, S.R. (2004). What is a hidden Markov model? *Nat. Biotechnol.* 22, 1315–1316.

Efron B., Hastie T., Johnstone I., and R. Tibshirani R. (2004). Least angle regression. *Annals of Statistics*, 32, pp. 407–499.

Efroymson, MA (1960) "Multiple regression analysis." In Ralston, A. and Wilf, HS, editors, *Mathematical Methods for Digital Computers*. Wiley.

Ellinghaus, D., Baurecht, H., Esparza-Gordillo, J., Rodríguez, E., Matanovic, A., Marenholz, I., Hübner, N., Schaarschmidt, H., Novak, N., Michel, S., et al. (2013). High-density genotyping study identifies four new susceptibility loci for atopic dermatitis. *Nat. Genet.* 45, 808–812.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636–640.

ENCODE Project Consortium, Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

- Ewens, W.J. (2004). *Mathematical population genetics* (New York: Springer).
- Excoffier, L., and Lischer, H.E.L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* *10*, 564–567.
- Excoffier, L., and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* *12*, 921–927.
- Federoff, M., Jimenez-Rolando, B., Nalls, M.A., and Singleton, A.B. (2012). A large study reveals no association between APOE and Parkinson's disease. *Neurobiol. Dis.* *46*, 389–392.
- Feussner, G., Feussner, V., Hoffmann, M.M., Lohrmann, J., Wieland, H., and März, W. (1998). Molecular basis of type III hyperlipoproteinemia in Germany. *Hum. Mutat.* *11*, 417–423.
- Fisher, R. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Trans R Soc Edinb* *52*, 399–433.
- Flister, M.J., Tsaih, S.-W., O'Meara, C.C., Endres, B., Hoffman, M.J., Geurts, A.M., Dwinell, M.R., Lazar, J., Jacob, H.J., and Moreno, C. (2013). Identifying multiple causative genes at a single GWAS locus. *Genome Res.* *23*, 1996–2002.
- Fojo, S.S., and Brewer, H.B. (1992). Hypertriglyceridaemia due to genetic defects in lipoprotein lipase and apolipoprotein C-II. *J. Intern. Med.* *231*, 669–677.
- Folsom, A.R. (2001). Hemostatic risk factors for atherothrombotic disease: an epidemiologic view. *Thromb. Haemost.* *86*, 366–373.
- Fowkes, F.G., Housley, E., Cawood, E.H., Macintyre, C.C., Ruckley, C.V., and Prescott, R.J. (1991). Edinburgh Artery Study: prevalence of asymptomatic and symptomatic peripheral arterial disease in the general population. *Int. J. Epidemiol.* *20*, 384–392.
- Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., et al. (2013). Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int. J. Epidemiol.* *42*, 97–110.
- Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R.B., Elliott, K.S., Lango, H., Rayner, N.W., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* *316*, 889–894.
- Galarneau, G., Palmer, C.D., Sankaran, V.G., Orkin, S.H., Hirschhorn, J.N., and Lettre, G. (2010). Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* *42*, 1049–1051.
- Gaunt, T.R., Rodriguez, S., Zapata, C., and Day, I.N.M. (2006). MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers. *BMC Bioinformatics* *7*, 227.
- Gaunt, T.R., Rodríguez, S., and Day, I.N. (2007). Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool "CubeX." *BMC*

Bioinformatics 8, 428.

Gaunt, T.R., Shah, S., Nelson, C.P., Drenos, F., Braund, P.S., Adeniran, I., Folkersen, L., Lawlor, D.A., Casas, J.-P., Amuzu, A., et al. (2012). Integration of genetics into a systems model of electrocardiographic traits using HumanCVD BeadChip. *Circ. Cardiovasc. Genet.* 5, 630–638.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korb, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 17, 669–681.

Ghebranious, N., Ivacic, L., Mallum, J., and Dokken, C. (2005). Detection of ApoE E2, E3 and E4 alleles using MALDI-TOF mass spectrometry and the homogeneous mass-extend technology. *Nucleic Acids Res.* 33, e149.

Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* 10, e1004383.

Global Lipids Genetics Consortium, Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283.

Golding, J., Pembrey, M., Jones, R., and ALSPAC Study Team (2001). ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr. Perinat. Epidemiol.* 15, 74–87.

Goldstein, J., Hobbs, H., and Brown, M. (1995). The metabolic bases of inherited disease. In *The metabolic bases of inherited disease* (New York: McGraw Hill).

Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 82, 100–112.

Gray, I.C., Campbell, D.A., and Spurr, N.K. (2000). Single nucleotide polymorphisms as tools in human genetics. *Hum. Mol. Genet.* 9, 2403–2408.

Grünhage, F., Acalovschi, M., Tirziu, S., Walier, M., Wienker, T.F., Ciocan, A., Mosteanu, O., Sauerbruch, T., and Lammert, F. (2007). Increased gallstone risk in humans conferred by common variant of hepatic ATP-binding cassette transporter for cholesterol. *Hepatol. Baltim. Md* 46, 793–801.

Grzegorzczak, M. (2010). An introduction to Gaussian Bayesian networks. *Methods Mol. Biol.* Clifton NJ 662, 121–147.

Hage, F.G., and Szalai, A.J. (2007). C-reactive protein gene polymorphisms, C-reactive protein blood levels, and cardiovascular disease risk. *J. Am. Coll. Cardiol.* 50, 1115–1122.

Han, T.S., Sattar, N., Williams, K., Gonzalez-Villalpando, C., Lean, M.E.J., and Haffner, S.M. (2002). Prospective study of C-reactive protein in relation to the development of diabetes and metabolic syndrome in the Mexico City Diabetes Study. *Diabetes Care* 25, 2016–2021.

- Hanukoglu, I. (1992). Steroidogenic enzymes: structure, function, and role in regulation of steroid hormone biosynthesis. *J. Steroid Biochem. Mol. Biol.* *43*, 779–804.
- Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox, M.J., Schneider, J.A., Moulin, D.S., and Clegg, J.B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* *60*, 772–789.
- Hartl, D.L. (2007). *Principles of population genetics* (Sunderland, Mass: Sinauer Associates).
- Hill, W.G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG Theor. Appl. Genet. Theor. Angew. Genet.* *38*, 226–231.
- Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* *106*, 9362–9367.
- Hingorani, A.D., and Psaty, B.M. (2009). Primary Prevention of Cardiovascular Disease: Time to Get More or Less Personal? *JAMA* *302*, 2144.
- Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* *44*, 955–959.
- Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* *5*, e1000529.
- Huang, X., Chen, P.C., and Poole, C. (2004). APOE-[epsilon]2 allele associated with higher prevalence of sporadic Parkinson disease. *Neurology* *62*, 2198–2202.
- Hunt, S.C., Kronenberg, F., Eckfeldt, J.H., Hopkins, P.N., Myers, R.H., and Heiss, G. (2001). Association of plasma bilirubin with coronary heart disease and segregation of bilirubin as a major gene trait: the NHLBI family heart study. *Atherosclerosis* *154*, 747–754.
- Hurst, J.W. (1998). Naming of the waves in the ECG, with a brief account of their genesis. *Circulation* *98*, 1937–1942.
- Inouye, M., Kettunen, J., Soininen, P., Silander, K., Ripatti, S., Kumpula, L.S., Hämäläinen, E., Jousilahti, P., Kangas, A.J., Männistö, S., et al. (2010). Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol. Syst. Biol.* *6*, 441.
- International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* *467*, 52–58.
- International HapMap Consortium (2003). The International HapMap Project. *Nature* *426*, 789–796.
- International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* *437*, 1299–1320.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945.

Ioannidis, J.P.A., Gwinn, M., Little, J., Higgins, J.P.T., Bernstein, J.L., Boffetta, P., Bondy, M., Bray, M.S., Brenchley, P.E., Buffler, P.A., et al. (2006). A road map for efficient and reliable human genome epidemiology. *Nat. Genet.* 38, 3–5.

Ioannidis, J.P.A., Patsopoulos, N.A., and Evangelou, E. (2007). Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* 2, e841.

Jaffrey, S.R., Snowman, A.M., Eliasson, M.J., Cohen, N.A., and Snyder, S.H. (1998). CAPON: a protein associated with neuronal nitric oxide synthase that regulates its interactions with PSD95. *Neuron* 20, 115–124.

Jeff, J.M., Ritchie, M.D., Denny, J.C., Kho, A.N., Ramirez, A.H., Crosslin, D., Armstrong, L., Basford, M.A., Wolf, W.A., Pacheco, J.A., et al. (2013). Generalization of Variants Identified by Genome-Wide Association Studies for Electrocardiographic Traits in African Americans: Generalization of ECG Associations. *Ann. Hum. Genet.* n/a–n/a.

Johansen, C.T., Wang, J., Lanktree, M.B., Cao, H., McIntyre, A.D., Ban, M.R., Martins, R.A., Kennedy, B.A., Hassell, R.G., Visser, M.E., et al. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* 42, 684–687.

Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J., and de Bakker, P.I.W. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinforma. Oxf. Engl.* 24, 2938–2939.

Johnston, D.E., and Kaplan, M.M. (1993). Pathogenesis and treatment of gallstones. *N. Engl. J. Med.* 328, 412–421.

Kalloo, A.N., and Kantsevoy, S.V. (2001). Gallstones and biliary disease. *Prim. Care* 28, 591–606, vii.

KANNEL, W.B., DAWBER, T.R., KAGAN, A., REVOTSKIE, N., and STOKES, J., 3rd (1961). Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. *Ann. Intern. Med.* 55, 33–50.

Kao, W.H.L., Arking, D.E., Post, W., Rea, T.D., Sotoodehnia, N., Prineas, R.J., Bishe, B., Doan, B.Q., Boerwinkle, E., Psaty, B.M., et al. (2009). Genetic variations in nitric oxide synthase 1 adaptor protein are associated with sudden cardiac death in US white community-based populations. *Circulation* 119, 940–951.

Kathiresan, S., Manning, A.K., Demissie, S., D'Agostino, R.B., Surti, A., Guiducci, C., Gianniny, L., Burt, N.P., Melander, O., Orho-Melander, M., et al. (2007). A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med. Genet.* 8 Suppl 1, S17.

Katsika, D., Magnusson, P., Krawczyk, M., Grünhage, F., Lichtenstein, P., Einarsson, C., Lammert, F., and Marschall, H.-U. (2010). Gallstone disease in Swedish twins: risk is associated with ABCG8 D19H genotype.

J. Intern. Med. 268, 279–285.

Kay, C., Skotte, N.H., Southwell, A.L., and Hayden, M.R. (2014). Personalized gene silencing therapeutics for Huntington disease. *Clin. Genet.*

Ke, X., Durrant, C., Morris, A.P., Hunt, S., Bentley, D.R., Deloukas, P., and Cardon, L.R. (2004). Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.* 13, 2557–2565.

Keating, B.J., Tischfield, S., Murray, S.S., Bhangale, T., Price, T.S., Glessner, J.T., Galver, L., Barrett, J.C., Grant, S.F.A., Farlow, D.N., et al. (2008). Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PloS One* 3, e3583.

Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.-P., Kangas, A.J., Soininen, P., Würtz, P., Silander, K., et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* 44, 269–276.

Kim, J.W., Hong, K.-W., Go, M.J., Kim, S.S., Tabara, Y., Kita, Y., Tanigawa, T., Cho, Y.S., Han, B.-G., and Oh, B. (2012). A Common Variant in SLC8A1 Is Associated with the Duration of the Electrocardiographic QT Interval. *Am. J. Hum. Genet.* 91, 180–184.

Kitsios, G.D., and Zintzaras, E. (2009). Genome-wide association studies: hypothesis-“free” or “engaged”? *Transl. Res. J. Lab. Clin. Med.* 154, 161–164.

Kline, K.G., Frewen, B., Bristow, M.R., Maccoss, M.J., and Wu, C.C. (2008). High quality catalog of proteotypic peptides from human heart. *J. Proteome Res.* 7, 5055–5061.

Klug, W.S. (2012). *Concepts of genetics* (San Francisco: Pearson Education).

Knowles, J.R. (1980). Enzyme-catalyzed phosphoryl transfer reactions. *Annu. Rev. Biochem.* 49, 877–919.

Knüppel, S., Esparza-Gordillo, J., Marenholz, I., Holzhütter, H.-G., Bauerfeind, A., Ruether, A., Weidinger, S., Lee, Y.-A., and Rohde, K. (2012). Multi-locus stepwise regression: a haplotype-based algorithm for finding genetic associations applied to atopic dermatitis. *BMC Med. Genet.* 13, 8.

Koboldt, D.C., Miller, R.D., and Kwok, P.-Y. (2006). Distribution of human SNPs and its effect on high-throughput genotyping. *Hum. Mutat.* 27, 249–254.

Koster, T., Blann, A.D., Briët, E., Vandenbroucke, J.P., and Rosendaal, F.R. (1995). Role of clotting factor VIII in effect of von Willebrand factor on occurrence of deep-vein thrombosis. *Lancet* 345, 152–155.

Krausz, C., and Chianese, C. (2014). Genetic testing and counselling for male infertility. *Curr. Opin. Endocrinol. Diabetes Obes.* 21, 244–250.

Kronenberg, F., Coon, H., Gutin, A., Abkevich, V., Samuels, M.E., Ballinger, D.G., Hopkins, P.N., and Hunt, S.C. (2002). A genome scan for loci influencing anti-atherogenic serum bilirubin levels. *Eur. J. Hum. Genet. EJHG* 10, 539–546.

Kuo, K.-K., Shin, S.-J., Chen, Z.-C., Yang, Y.-H.C., Yang, J.-F., and Hsiao, P.-J. (2008). Significant association of ABCG5 604Q and ABCG8 D19H polymorphisms with gallstone disease. *Br. J. Surg.* 95, 1005–1011.

Kutner, M. H.; Nachtsheim, C. J.; Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill Irwin.

Kuulasmaa, K., Tunstall-Pedoe, H., Dobson, A., Fortmann, S., Sans, S., Tolonen, H., Evans, A., Ferrario, M., and Tuomilehto, J. (2000). Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA Project populations. *Lancet* 355, 675–687.

Laaksonen, D.E., Niskanen, L., Nyyssönen, K., Punnonen, K., Tuomainen, T.-P., Valkonen, V.-P., Salonen, R., and Salonen, J.T. (2004). C-reactive protein and the development of the metabolic syndrome and diabetes in middle-aged men. *Diabetologia* 47, 1403–1410.

Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* 45, 1452–1458.

Lander, E.S. (1996). The new genomics: global views of biology. *Science* 274, 536–539.

De Lange, M., Snieder, H., Ariëns, R.A., Spector, T.D., and Grant, P.J. (2001). The genetics of haemostasis: a twin study. *Lancet* 357, 101–105.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–838.

Law, M.R., Wald, N.J., and Rudnicka, A.R. (2003). Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *BMJ* 326, 1423.

Lawlor, D.A., Bedford, C., Taylor, M., and Ebrahim, S. (2003). Geographical variation in cardiovascular disease, risk factors, and their control in older women: British Women's Heart and Health Study. *J. Epidemiol. Community Health* 57, 134–140.

Levene, P. (1919). The Structure of Yeast Nucleic Acid Iv. Ammonia Hydrolysis. *J Biol Chem* 40, 415–424.

Lewontin, R.C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* 49, 49–67.

Li JH, Das K, Fu G, Li RZ, Wu RL. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatics* 27 (4): 516-523.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* 10, 387–406.

Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34, 816–834.

- Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255–268.
- Lin, L., and Torbeck, L.D. (1998). Coefficient of accuracy and concordance correlation coefficient: new statistics for methods comparison. *PDA J. Pharm. Sci. Technol.* PDA 52, 55–59.
- Lin, J.-P., Cupples, L.A., Wilson, P.W.F., Heard-Costa, N., and O'Donnell, C.J. (2003). Evidence for a gene influencing serum bilirubin on chromosome 2q telomere: a genomewide scan in the Framingham study. *Am. J. Hum. Genet.* 72, 1029–1034.
- Liu, J.-S., Huang, C.-S., and Lien, H.-H. (2002). Structural analysis of gallstones with thin-section petrographic microscopy: a study of 100 gallstones from Taiwanese patients. *J. Lab. Clin. Med.* 140, 387–390.
- Lloyd-Jones, D., Adams, R.J., Brown, T.M., Carnethon, M., Dai, S., De Simone, G., Ferguson, T.B., Ford, E., Furie, K., Gillespie, C., et al. (2010). Heart disease and stroke statistics--2010 update: a report from the American Heart Association. *Circulation* 121, e46–e215.
- Lloyd-Jones, D.M., Nam, B.-H., D'Agostino, R.B., Sr, Levy, D., Murabito, J.M., Wang, T.J., Wilson, P.W.F., and O'Donnell, C.J. (2004). Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. *JAMA J. Am. Med. Assoc.* 291, 2204–2211.
- Loos, R.J.F., Lindgren, C.M., Li, S., Wheeler, E., Zhao, J.H., Prokopenko, I., Inouye, M., Freathy, R.M., Attwood, A.P., Beckmann, J.S., et al. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat. Genet.* 40, 768–775.
- Ma, L., Han, S., Yang, J., and Da, Y. (2010). Multi-locus Test Conditional on Confirmed Effects Leads to Increased Power in Genome-wide Association Studies. *PLoS ONE* 5, e15006.
- Madsen, R., Lundstedt, T., and Trygg, J. (2010). Chemometrics in metabolomics—A review in human disease diagnosis. *Anal. Chim. Acta* 659, 23–33.
- Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* 456, 18–21.
- Maher, B. (2012). ENCODE: The human encyclopaedia. *Nature* 489, 46–48.
- Mahley, R.W. (1988). Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. *Science* 240, 622–630.
- Mallows, C.L. (1973). Some Comments on C P. *Technometrics* 15, 661.
- Mann, D. (2008). *Pathophysiology of Heart Failur* (SAUNDERS ELSEVIER).
- Manolescu, A.R., Augustin, R., Moley, K., and Cheeseman, C. (2007). A highly conserved hydrophobic motif in the exofacial vestibule of fructose transporting SLC2A proteins acts as a critical determinant of their substrate selectivity. *Mol. Membr. Biol.* 24, 455–463.

Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363, 166–176.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753.

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., et al. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* 78, 437–450.

Mardis, E.R. (2006). Anticipating the 1,000 dollar genome. *Genome Biol.* 7, 112.

Marenberg, M.E., Risch, N., Berkman, L.F., Floderus, B., and de Faire, U. (1994). Genetic susceptibility to death from coronary heart disease in a study of twins. *N. Engl. J. Med.* 330, 1041–1046.

Marin-Valencia, I., Good, L.B., Ma, Q., Malloy, C.R., and Pascual, J.M. (2013). Heptanoate as a neural fuel: energetic and neurotransmitter precursors in normal and glucose transporter I-deficient (G1D) brain. *J. Cereb. Blood Flow Metab. Off. J. Int. Soc. Cereb. Blood Flow Metab.* 33, 175–182.

Marmot, M., Banks, J., Blundell, R., Lessof, C., and Nazroo, J. (2003). *Health, Wealth and Lifestyles of the Older Population in England: The 2002 English Longitudinal Study of Ageing*. London: Institute for Fiscal Studies.

Marmot, M.G., Smith, G.D., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E., and Feeney, A. (1991). Health inequalities among British civil servants: the Whitehall II study. *Lancet* 337, 1387–1393.

Marroni, F., Pfeufer, A., Aulchenko, Y.S., Franklin, C.S., Isaacs, A., Pichler, I., Wild, S.H., Oostra, B.A., Wright, A.F., Campbell, H., et al. (2009). A genome-wide association scan of RR and QT interval duration in 3 European genetically isolated populations: the EUROSPAN project. *Circ. Cardiovasc. Genet.* 2, 322–328.

McCarthy, M.I., and Hirschhorn, J.N. (2008). Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.* 17, R156–165.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26, 2069–2070.

McVean, G.A.T., and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 360, 1387–1393.

MENDEL, G. (1950). Gregor Mendel's letters to Carl Nägeli, 1866-1873. *Genetics* 35, 1–29.

Miller, N.E., and Miller, G.J. (1975). Letter: High-density lipoprotein and atherosclerosis. *Lancet* 1, 1033.

Miller, M., Rhyne, J., Hamlette, S., Birnbaum, J., and Rodriguez, A. (2003). Genetics of HDL regulation in humans. *Curr. Opin. Lipidol.* 14, 273–279.

De Miranda Chagas, S.V., Kanaan, S., Chung Kang, H., Cagy, M., de Abreu, R.E., da Silva, L.A., Garcia, R.C., and Garcia Rosa, M.L. (2011). Environmental factors, familial aggregation and heritability of total cholesterol, low density lipoprotein-cholesterol and high density lipoprotein-cholesterol in a Brazilian population assisted by the Family Doctor Program. *Public Health* 125, 329–337.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–777.

Montgomery, S.B., Goode, D.L., Kvikstad, E., Albers, C.A., Zhang, Z.D., Mu, X.J., Ananda, G., Howie, B., Karczewski, K.J., Smith, K.S., et al. (2013). The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* 23, 749–761.

Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* 34, 188–193.

Motyer JA, McKendry C, Galbraith S and Wilson SR. (2011). LASSO model selection with post-processing for a genome-wide association study data set. *BMC Proc.* 2011; 5(Suppl 9): S24.

Mukhopadhyay, I., Feingold, E., Weeks, D.E., and Thalamuthu, A. (2010). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet. Epidemiol.* 34, 213–221.

Murabito, J.M., Pencina, M.J., Nam, B.-H., D'Agostino, R.B., Sr, Wang, T.J., Lloyd-Jones, D., Wilson, P.W.F., and O'Donnell, C.J. (2005). Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. *JAMA J. Am. Med. Assoc.* 294, 3117–3123.

Musumeci, L., Arthur, J.W., Cheung, F.S.G., Hoque, A., Lippman, S., and Reichardt, J.K.V. (2010). Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum. Mutat.* 31, 67–73.

Musunuru, K., Lettre, G., Young, T., Farlow, D.N., Pirruccello, J.P., Ejebe, K.G., Keating, B.J., Yang, Q., Chen, M.-H., Lapchuk, N., et al. (2010a). Candidate gene association resource (CARE): design, methods, and proof of concept. *Circ. Cardiovasc. Genet.* 3, 267–275.

Musunuru, K., Pirruccello, J.P., Do, R., Peloso, G.M., Guiducci, C., Sougnez, C., Garimella, K.V., Fisher, S., Abreu, J., Barry, A.J., et al. (2010b). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N. Engl. J. Med.* 363, 2220–2227.

Namboodiri, K.K., Kaplan, E.B., Heuch, I., Elston, R.C., Green, P.P., Rao, D.C., Laskarzewski, P., Glueck, C.J., and Rifkind, B.M. (1985). The Collaborative Lipid Research Clinics Family Study: biological and cultural determinants of familial resemblance for plasma lipids and lipoproteins. *Genet. Epidemiol.* 2, 227–254.

Newton-Cheh, C., Eijgelsheim, M., Rice, K.M., de Bakker, P.I.W., Yin, X., Estrada, K., Bis, J.C., Marciante, K.,

Rivadeneira, F., Noseworthy, P.A., et al. (2009). Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nat. Genet.* **41**, 399–406.

Nolte, I.M., Wallace, C., Newhouse, S.J., Waggott, D., Fu, J., Soranzo, N., Gwilliam, R., Deloukas, P., Savelieva, I., Zheng, D., et al. (2009). Common Genetic Variation Near the Phospholamban Gene Is Associated with Cardiac Repolarisation: Meta-Analysis of Three Genome-Wide Association Studies. *PLoS ONE* **4**, e6138.

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M., and Franke, A. (2009). A comprehensive evaluation of SNP genotype imputation. *Hum. Genet.* **125**, 163–171.

Novotný, L., and Vitek, L. (2003). Inverse relationship between serum bilirubin and atherosclerosis in men: a meta-analysis of published studies. *Exp. Biol. Med.* Maywood NJ **228**, 568–571.

Nyholt, D.R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.* **74**, 765–769.

Ogden, C.L., Carroll, M.D., Curtin, L.R., McDowell, M.A., Tabak, C.J., and Flegal, K.M. (2006). Prevalence of overweight and obesity in the United States, 1999–2004. *JAMA J. Am. Med. Assoc.* **295**, 1549–1555.

Palmer, C.N.A., Irvine, A.D., Terron-Kwiatkowski, A., Zhao, Y., Liao, H., Lee, S.P., Goudie, D.R., Sandilands, A., Campbell, L.E., Smith, F.J.D., et al. (2006). Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat. Genet.* **38**, 441–446.

Park, J.-H., Gail, M.H., Weinberg, C.R., Carroll, R.J., Chung, C.C., Wang, Z., Chanock, S.J., Fraumeni, J.F., Jr, and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18026–18031.

Paternoster, L., Standl, M., Chen, C.-M., Ramasamy, A., Bønnelykke, K., Duijts, L., Ferreira, M.A., Alves, A.C., Thyssen, J.P., Albrecht, E., et al. (2012). Meta-analysis of genome-wide association studies identifies three new risk loci for atopic dermatitis. *Nat. Genet.* **44**, 187–192.

Pelak, K., Shianna, K.V., Ge, D., Maia, J.M., Zhu, M., Smith, J.P., Cirulli, E.T., Fellay, J., Dickson, S.P., Gumbs, C.E., et al. (2010). The characterization of twenty sequenced human genomes. *PLoS Genet.* **6**, e1001111.

Pereira, T.V., Patsopoulos, N.A., Salanti, G., and Ioannidis, J.P.A. (2009). Discovery properties of genome-wide association signals from cumulatively combined data sets. *Am. J. Epidemiol.* **170**, 1197–1206.

Pfeiffer, R., Mitchell, H., and Pee, D. (2009). On Combining Data From Genome-Wide Association Studies to Discover Disease-Associated SNPs. *Stat. Sci.* **24**, 547–560.

Pfeufer, A., Sanna, S., Arking, D.E., Müller, M., Gateva, V., Fuchsberger, C., Ehret, G.B., Orrú, M., Pattaro, C., Köttgen, A., et al. (2009). Common variants at ten loci modulate the QT interval duration in the QTSCD Study. *Nat. Genet.* **41**, 407–414.

Pfeufer, A., van Noord, C., Marciante, K.D., Arking, D.E., Larson, M.G., Smith, A.V., Tarasov, K.V., Müller, M.,

- Sotoodehnia, N., Sinner, M.F., et al. (2010). Genome-wide association study of PR interval. *Nat. Genet.* **42**, 153–159.
- Pilia, G., Chen, W.-M., Scuteri, A., Orrú, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* **2**, e132.
- Plomin, R., Haworth, C.M.A., and Davis, O.S.P. (2009). Common disorders are quantitative traits. *Nat. Rev. Genet.* **10**, 872–878.
- Pollin, T.I., Hsueh, W.-C., Steinle, N.I., Snitker, S., Shuldiner, A.R., and Mitchell, B.D. (2004). A genome-wide scan of serum lipid levels in the Old Order Amish. *Atherosclerosis* **173**, 89–96.
- Pradhan, A.D., Manson, J.E., Rifai, N., Buring, J.E., and Ridker, P.M. (2001). C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA J. Am. Med. Assoc.* **286**, 327–334.
- Price, J.F., Reynolds, R.M., Mitchell, R.J., Williamson, R.M., Fowkes, F.G.R., Deary, I.J., Lee, A.J., Frier, B.M., Hayes, P.C., and Strachan, M.W.J. (2008). The Edinburgh Type 2 Diabetes Study: study protocol. *BMC Endocr. Disord.* **8**, 18.
- Prontera, P., Bartocci, A., Ottaviani, V., Isidori, I., Rogaia, D., Ardisia, C., Guercini, G., Mencarelli, A., and Dotti, E. (2013). Aicardi syndrome associated with autosomal genomic imbalance: coincidence or evidence for autosomal inheritance with sex-limited expression? *Mol. Syndromol.* **4**, 197–202.
- Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43**, 977–983.
- Pulst, S.M. (1999). Genetic linkage analysis. *Arch. Neurol.* **56**, 667–672.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575.
- Qin, Z.S., Niu, T., and Liu, J.S. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **71**, 1242–1247.
- Race, Ethnicity, and Genetics Working Group (2005). The use of racial, ethnic, and ancestral categories in human genetics research. *Am. J. Hum. Genet.* **77**, 519–532.
- Rader, D.J., Cohen, J., and Hobbs, H.H. (2003). Monogenic hypercholesterolemia: new insights in pathogenesis and treatment. *J. Clin. Invest.* **111**, 1795–1803.
- Rahim, N.G., Harismendy, O., Topol, E.J., and Frazer, K.A. (2008). Genetic determinants of phenotypic diversity in humans. *Genome Biol.* **9**, 215.
- Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends Genet. TIG* **17**, 502–510.

- Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. *Nature* 411, 199–204.
- Remme, C.A., Wilde, A.A.M., and Bezzina, C.R. (2008). Cardiac sodium channel overlap syndromes: different faces of SCN5A mutations. *Trends Cardiovasc. Med.* 18, 78–87.
- Richardson, T.G., Thomas, E.C., Sessions, R.B., Lawlor, D.A., Tavaré, J.M., and Day, I.N.M. (2013). Structural and population-based evaluations of TBC1D1 p.Arg125Trp. *PloS One* 8, e63897.
- Ridker, P.M., Cushman, M., Stampfer, M.J., Tracy, R.P., and Hennekens, C.H. (1997). Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *N. Engl. J. Med.* 336, 973–979.
- Ridker, P.M., Rifai, N., Rose, L., Buring, J.E., and Cook, N.R. (2002). Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. *N. Engl. J. Med.* 347, 1557–1565.
- Rieder, M.J., Taylor, S.L., Clark, A.G., and Nickerson, D.A. (1999). Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.* 22, 59–62.
- Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* 273, 1516–1517.
- Robinson, W.P., Asmussen, M.A., and Thomson, G. (1991). Three-locus systems impose additional constraints on pairwise disequilibria. *Genetics* 129, 925–930.
- Rodriguez, S., Steer, C.D., Farrow, A., Golding, J., and Day, I.N.M. (2013). Dependence of deodorant usage on ABCC11 genotype: scope for personalized genetics in personal hygiene. *J. Invest. Dermatol.* 133, 1760–1767.
- Rodriguez, S., Gaunt, T., Guo, Y., Zheng, J., Tang, W., Danish, F., Johnson, A., Castillo, B., Li, Y., and Hakonarson, H. (2014). Lipids, obesity and gallbladder disease in women: insights from genetic studies using the cardiovascular gene-centric 50K SNP array.
- Rodríguez, E., Baurecht, H., Herberich, E., Wagenpfeil, S., Brown, S.J., Cordell, H.J., Irvine, A.D., and Weidinger, S. (2009). Meta-analysis of filaggrin polymorphisms in eczema and asthma: robust risk factors in atopic disease. *J. Allergy Clin. Immunol.* 123, 1361–1370.e7.
- Roger, V.L., Go, A.S., Lloyd-Jones, D.M., Benjamin, E.J., Berry, J.D., Borden, W.B., Bravata, D.M., Dai, S., Ford, E.S., Fox, C.S., et al. (2012). Heart disease and stroke statistics--2012 update: a report from the American Heart Association. *Circulation* 125, e2–e220.
- Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837.
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265, 687–695.

- Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H.M., Jackson, A.U., Piras, M.G., Usala, G., Maninchedda, G., Sassu, A., et al. (2011). Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLoS Genet.* 7, e1002198.
- Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Res.* 22, 1748–1759.
- Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644.
- Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* 43, 969–976.
- Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Stat.* 6, 461–464.
- Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830.
- Seminara, D., Khoury, M.J., O’Brien, T.R., Manolio, T., Gwinn, M.L., Little, J., Higgins, J.P.T., Bernstein, J.L., Boffetta, P., Bondy, M., et al. (2007). The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiol. Camb. Mass* 18, 1–8.
- Shaffer, E.A. (2006). Gallstone disease: Epidemiology of gallbladder stone disease. *Best Pract. Res. Clin. Gastroenterol.* 20, 981–996.
- Shah, S., Casas, J.P., Gaunt, T.R., Cooper, J., Drenos, F., Zabaneh, D., Swerdlow, D.I., Shah, T., Sofat, R., Palmen, J., et al. (2013a). Influence of common genetic variation on blood lipid levels, cardiovascular risk, and coronary events in two British prospective cohort studies. *Eur. Heart J.* 34, 972–981.
- Shah, T., Engmann, J., Dale, C., Shah, S., White, J., Giambartolomei, C., McLachlan, S., Zabaneh, D., Cavadino, A., Finan, C., et al. (2013b). Population genomics of cardiometabolic traits: design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PloS One* 8, e71345.
- Shaper, A.G., Pocock, S.J., Walker, M., Cohen, N.M., Wale, C.J., and Thomson, A.G. (1981). British Regional Heart Study: cardiovascular risk factors in middle-aged men in 24 towns. *Br. Med. J. Clin. Res. Ed* 283, 179–186.
- Sherry, S.T., Ward, M., and Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* 9, 677–679.
- Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311.
- Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L.A., Edwards, K.J., Day, I.N.M., and Gaunt,

T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 34, 57–65.

Sirtori, C.R., and Fumagalli, R. (2006). LDL-cholesterol lowering or HDL-cholesterol raising for cardiovascular prevention. A lesson from cholesterol turnover studies and others. *Atherosclerosis* 186, 1–11.

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477–485.

Smith, G.D., and Ebrahim, S. (2003). "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32, 1–22.

Smith, A.J.P., Palmen, J., Putt, W., Talmud, P.J., Humphries, S.E., and Drenos, F. (2010a). Application of statistical and functional methodologies for the investigation of genetic determinants of coronary heart disease biomarkers: lipoprotein lipase genotype and plasma triglycerides as an exemplar. *Hum. Mol. Genet.* 19, 3936–3947.

Smith, J.G., Magnani, J.W., Palmer, C., Meng, Y.A., Soliman, E.Z., Musani, S.K., Kerr, K.F., Schnabel, R.B., Lubitz, S.A., Sotoodehnia, N., et al. (2011). Genome-Wide Association Studies of the PR Interval in African Americans. *PLoS Genet.* 7, e1001304.

Smith, J.G., Avery, C.L., Evans, D.S., Nalls, M.A., Meng, Y.A., Smith, E.N., Palmer, C., Tanaka, T., Mehra, R., Butler, A.M., et al. (2012). Impact of Ancestry and Common Genetic Variants on QT Interval in African Americans. *Circ. Cardiovasc. Genet.* 5, 647–655.

Smith, N.L., Chen, M.-H., Dehghan, A., Strachan, D.P., Basu, S., Soranzo, N., Hayward, C., Rudan, I., Sabater-Lleal, M., Bis, J.C., et al. (2010b). Novel Associations of Multiple Genetic Loci With Plasma Levels of Factor VII, Factor VIII, and von Willebrand Factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation* 121, 1382–1392.

Soininen, P., Kangas, A.J., Würtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., Järvelin, M.-R., Kähönen, M., Lehtimäki, T., Viikari, J., et al. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *The Analyst* 134, 1781–1785.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42, 937–948.

Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52, 506–516.

StataCorp (2013). Stata Statistical Software: Release 13.

Stephens, M., and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* 76, 449–462.

Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978–989.

Storz, G. (2002). An expanding universe of noncoding RNAs. *Science* 296, 1260–1263.

Strachan, T., and Read, A. (2011). *Human molecular genetics* (New York: Garland Science).

Strittmatter, W.J., Saunders, A.M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G.S., and Roses, A.D. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl. Acad. Sci. U. S. A.* 90, 1977–1981.

Südhof, T.C., Goldstein, J.L., Brown, M.S., and Russell, D.W. (1985). The LDL receptor gene: a mosaic of exons shared with different proteins. *Science* 228, 815–822.

Tall, A.R., Yvan-Charvet, L., and Wang, N. (2007). The failure of torcetrapib: was it the molecule or the mechanism? *Arterioscler. Thromb. Vasc. Biol.* 27, 257–260.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707–713.

Tibshirani, R., (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B, Volume 58, Issue 1*, 267–288

The ARIC investigators (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am. J. Epidemiol.* 129, 687–702.

The Women’s Health Initiative Study Group (1998). Design of the Women’s Health Initiative clinical trial and observational study. *Control. Clin. Trials* 19, 61–109.

Thompson, J.R., Gögele, M., Weichenberger, C.X., Modenese, M., Attia, J., Barrett, J.H., Boehnke, M., De Grandi, A., Domingues, F.S., Hicks, A.A., et al. (2013). SNP prioritization using a Bayesian probability of association. *Genet. Epidemiol.* 37, 214–221.

Thorleifsson, G., Walters, G.B., Gudbjartsson, D.F., Steinthorsdottir, V., Sulem, P., Helgadóttir, A., Styrkarsdóttir, U., Gretarsdóttir, S., Thorlacius, S., Jonsdóttir, I., et al. (2009). Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat. Genet.* 41, 18–24.

Toth, P.P. (2005). Cardiology patient page. The “good cholesterol”: high-density lipoprotein. *Circulation* 111, e89–91.

Tregouet, D.A., Escolano, S., Tiret, L., Mallet, A., and Golmard, J.L. (2004). A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm. *Ann. Hum. Genet.* 68, 165–177.

Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G., et al. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* 43, 1193–1201.

Tukiainen, T., Kettunen, J., Kangas, A.J., Lyytikäinen, L.-P., Soininen, P., Sarin, A.-P., Tikkanen, E., O'Reilly, P.F., Savolainen, M.J., Kaski, K., et al. (2012). Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Hum. Mol. Genet.* 21, 1444–1455.

Valdar, W., Sabourin, J., Nobel, A., and Holmes, C.C. (2012). Reprioritizing genetic associations in hit regions using LASSO-based resample model averaging. *Genet. Epidemiol.* 36, 451–462.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8, e1002793.

Wadsworth, M., Kuh, D., Richards, M., and Hardy, R. (2006). Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *Int. J. Epidemiol.* 35, 49–54.

Wang, D.Q.-H., and Afdhal, N.H. (2004). Genetic analysis of cholesterol gallstone formation: searching for Lith (gallstone) genes. *Curr. Gastroenterol. Rep.* 6, 140–150.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* 14, 125–138.

Wellcome Trust Case Control Consortium, Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M.M., Auton, A., Myers, S., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* 44, 1294–1301.

Wessel, J., and Schork, N.J. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.* 79, 792–806.

Westaway, S.K., Reinier, K., Huertas-Vazquez, A., Evanado, A., Teodorescu, C., Navarro, J., Sinner, M.F., Gunson, K., Jui, J., Spooner, P., et al. (2011). Common variants in CASQ2, GPD1L, and NOS1AP are significantly associated with risk of sudden death in patients with coronary artery disease. *Circ. Cardiovasc. Genet.* 4, 397–402.

WHO (2005). Preventing chronic diseases: A vital investment.

WHO (2009). World health statistics 2009. Geneva: World Health Organization.

Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinform. Oxf. Engl.* 26, 2190–2191.

Wiltshire, S., Morris, A.P., and Zeggini, E. (2008). Examining the statistical properties of fine-scale mapping in large-scale association studies. *Genet. Epidemiol.* 32, 204–214.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., et al. (2007). HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35, D521–526.

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., et al. (2013). HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res.* *41*, D801–807.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* *42*, 565–569.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* *88*, 76–82.

Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* *44*, 369–375, S1–3.

Yoshiura, K., Kinoshita, A., Ishida, T., Ninokata, A., Ishikawa, T., Kaname, T., Bannai, M., Tokunaga, K., Sonoda, S., Komaki, R., et al. (2006). A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat. Genet.* *38*, 324–330.

Wu TT, Chen YF, Hastie T, Sobel E and Lange K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25 (6): 714-721.

Ypma, T.J. (1995). Historical Development of the Newton–Raphson Method. *SIAM Rev.* *37*, 531–551.

Yuan, Y., Xiang (1999). Step-sizes for the gradient method. *AMSIP Stud. Adv. Math.* *42*, 758.

Zeggini, E., and Ioannidis, J.P.A. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics* *10*, 191–201.

Zhang, W., Collins, A., Gibson, J., Tapper, W.J., Hunt, S., Deloukas, P., Bentley, D.R., and Morton, N.E. (2004). Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl. Acad. Sci. U. S. A.* *101*, 18075–18080.

Zheng, J., Gaunt, T.R., and Day, I.N.M. (2013). Sequential sentinel SNP Regional Association Plots (SSS-RAP): an approach for testing independence of SNP association signals using meta-analysis data. *Ann. Hum. Genet.* *77*, 67–79.

Zheng, J., Rodriguez, S., White, J., Giambartolomei, C., Zabaneh, D., Morris, R., Kumari, M., Casas, J., Hingorani, A., UCLEB, et al. (2015). Haplotype-based genetic fine mapping using complex traits meta-analysis summary results. *Bioinformatics*. *Reveiwed*.

Zimmermann, K., Leffler, A., Babes, A., Cendan, C.M., Carr, R.W., Kobayashi, J., Nau, C., Wood, J.N., and Reeh, P.W. (2007). Sensory neuron sodium channel Nav1.8 is essential for pain at low temperatures. *Nature* *447*, 855–858.

Zou H. and Hastie T. (2005). Regularization and variable selection via the elastic net. *JRST*, 67, pp. 301–

320.

CHAPTER 8. APPENDIX

8.1. Supplementary Notes

8.1.1. Codes for LiftOver

The code to run liftOver is as follows:

```
liftOver input.bed hg18ToHg19.over.chain.gz output.bed unlifted.bed
```

Where input.bed is the input file, hg18ToHg19.over.chain.gz is the chain file and output.bed is the output file with SNP information after lift over.

8.1.2. R codes used for model selection methods

Firstly, I created a data frame called “dataframe”, which contain genotype and phenotype information. Then I created a linear model called “linear” using the following code:

```
linear <- lm(phenotype~., dataframe)
```

The R packages I used for the first four model selection approaches are “leap” and “bestglm”. The R codes are:

1. AIC: `bestglm(dataframe, IC="AIC")`
2. BIC: `bestglm(dataframe, IC="BIC")`
3. Cross Validation: `bestglm(dataframe, IC="CV")`
4. LOOCV: `bestglm(dataframe, IC="LOOCV")`

The R package I used for the next two methods is called “leap”. I created a regression model using the following code:

```
subset <- regsubsets(y~., dataframe, nvmax=snpno., really.big=T)
```

5. Mallows C_p : `plot(subset, scale="Cp")`
6. Adjusted r^2 : `plot(subset, scale="adjr2")`

For the next three methods, I used the basic R code "step"

7. StepwiseAIC: `step(linear, dataframe, direction="both", k=2, trace=0)`
8. stepwiseBIC: `step(linear, dataframe, direction="both", k=log(nrow(dataframe)), trace=0)`
9. stepwise C_p : `step(linear, dataframe, direction="both", scale=(summary(linear)$sigma)^z, trace=0)`

For the last methods, I used the basic linear model "lm"

10. stepwise regression with p value threshold = 0.05:

```
model_0 <- lm (phenotype~., dataframe)

summary(model_0) and knock out the SNP with highest p value (i.e. snp1)

model_1 <- update(model_0, ~.-snp1)

summary(model_1)

and so on
```

Until all reminding SNPs in the model reach the p value threshold of 0.05

8.1.3. Funding section of the samples

BWHHS

BWHHS (British Women's Heart and Health Study). BWHHS is supported by funding from the British Heart Foundation and the Department of Health Policy Research Programme

(England). We thank the BWHHS data collection team, General Practitioners who helped with recruitment of participants and the participants. We thank all of the participants and the general practitioners, research nurses, and data management staff who supported data collection and preparation. The BWHHS is coordinated by Shah Ebrahim (PI), D.L., and J.-P.C., with genotyping funded by the BHF (PG/07/131/24254, PI T.G.). This work was carried out in a Centre that receives core funding from the UK Medical Research Council (G0600705) and the University of Bristol.

HapMap

The USA National Institutes of Health, the National Human Genome Research Institute, the National Institute on Deafness and Other Communication Disorders and the Wellcome Trust for supporting the majority of this work. Funding was also provided by the Louis-Jeantet Foundation and the NCCR 'Frontiers in Genetics' (Swiss National Science Foundation).

1000 Genomes

This research was supported in part by Wellcome Trust grants WT098051 to R.M.D., M.E.H. and C.T.S.; WT090532/Z/09/Z, WT085475/Z/08/Z and WT095552/Z/11/Z to P.Do.; WT086084/Z/08/Z and WT090532/Z/09/Z to G.A.M.; WT089250/Z/09/Z to I.M.; WT085532AIA to P.F.; Medical Research Council grant G0900747(91070) to G.A.M.; British Heart Foundation grant RG/09/12/28096 to C.A.A.; the National Basic Research Program of China (973 program no.2011CB809201, 2011CB809202 and 2011CB809203); the Chinese 863 program (2012AA02A201); the National Natural Science Foundation of China (30890032,31161130357); the Shenzhen Key Laboratory of Transomics Biotechnologies

(CXB201108250096A); the Shenzhen Municipal Government of China (grants ZYC200903240080A and ZYC201105170397A); Guangdong Innovative Research Team Program (no. 2009010016); BMBF grant 01GS08201 to H.Le.; BMBF grant 0315428A to R.H.; the Max Planck Society; Swiss National Science Foundation 31003A_130342 to E.T.D.; Swiss National Science Foundation NCCR 'Frontiers in Genetics' grant to E.T.D.; Louis Jeantet Foundation grant to E.T.D.; Biotechnology and Biological Sciences Research Council (BBSRC) grantBB/I021213/1 to A.R.-L.; German Research Foundation (Emmy Noether Fellowship KO 4037/1-1) to J.O.K.; Netherlands Organization for Scientific Research VENI grant 639.021.125 to K.Y.; Beatriu de Pinos Program grants 2006BP-A 10144 and 2009BP-B 00274 to M.V.; Israeli Science Foundation grant 04514831 to E.H.; Genome Que'bec and the Ministry of Economic Development, Innovation and Trade grant PSR-SIIRI-195 to P.Aw.; National Institutes of Health (NIH) grants UO1HG5214, RC2HG5581 and R01MH84698 to G.R.A.; R01HG4719 and R01HG3698 to G.T.M.; RC2HG5552 and UO1HG6513 to G.R.A. and G.T.M.; R01HG4960 and R01HG5701 to B.L.B.; U01HG5715 to C.D.B. and A.G.C.; 32GM8283 to D.Cl.; U01HG5208 to M.J.D.; U01HG6569 to M.A.D.; R01HG2898 and R01CA166661 to .E.D.; UO1HG5209, UO1HG5725 and P41HG4221 to C.Le.; P01HG4120 to E.E.E.; U01HG5728 to Yu.F.; U54HG3273 and U01HG5211 to R.A.G.; R01HL95045 to S.B.G.; U41HG4568 to S.J.K.; P41HG2371 to W.J.K.; ES015794, AI077439, HL088133 and HL078885 to E.G.B.; RC2HL102925 to S.B.G. and D.M.A.; R01GM59290 to L.B.J. and M.A.B.; U54HG3067 to E.S.L. and S.B.G.; T15LM7033 to B.K.M.; T32HL94284 to J.L.R.-F.; DP2OD6514 and BAA-NIAID-DAIT-NIHAI2009061 to P.C.S.; T32GM7748 to X.S.; U54HG3079 to R.K.W.; UL1RR024131 to R.D.H.; HHSN268201100040C to the Coriell Institute for Medical Research;

a Sandler Foundation award and an American Asthma Foundation award to E.G.B.; an IBM Open Collaborative Research Program award to Y.B.; an A.G. Leventis Foundation scholarship to D.K.X.; a Wolfson Royal Society Merit Award to P.Do.; a Howard Hughes Medical Institute International Fellowship award to P.H.S.; a grant from T. and V. Stanley to S.C.Y.; and a Mary Beryl Patch Turnbull Scholar Program award to K.C.B. E.H. is a faculty fellow of the Edmond J. Safra Bioinformatics program at Tel-Aviv University. E.E.E. and D.H. are investigators of the Howard Hughes Medical Institute. M.V.G. is a long-term fellow of EMBO.

ALSPAC

The UK Medical Research Council (Grant ref:74882), the Wellcome Trust (Grant ref: 076467), and the University of Bristol provide core support for the Avon Longitudinal Study of Parents and their Children (ALSPAC). The Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and 23andMe generated the ALSPAC GWA data. The Wellcome Trust and Swiss National Science Foundation funded the expression data.

UCLEB

UCLEB (University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol). The UCLEB consortium is funded by a British Heart Foundation programme grant (ref RG/10/12/28456).

CARe

The following parent studies, funded by the listed National Institutes of Health grants, have contributed parent study data, ancillary study data, and DNA samples through the Broad Institute (N01-HC-65226) to create this genotype/phenotype database for wide dissemination to the biomedical research community: Atherosclerotic Risk in Communities: University of North Carolina at Chapel Hill (N01-HC-55015, N01-HC-55018).

GIANT

US National Institutes of Health grants CA65725, CA87969, CA49449, CA67262, CA50385, DK062370, DK072193, DK075787, HG02651, HL084729, HL087679 (through STAMPEED, 1RL1MH083268), 5UO1CA098233, 1Z01 HG000024, 1RL1MH083268, T32 DK07191, F32 DK079466, K23 DK080145, K23 DK067288, CIDR NIH Contract Number N01-HG-65403, NIA contract NO1-AG-1-2109; the Intramural Research Program of the Division of Cancer Epidemiology and Genetics; contracts from the Division of Cancer Prevention, National Cancer Institute and EU FP6 funding (contract no LSHM-CT-2003-503041); GlaxoSmithKline; the Faculty of Biology and Medicine of Lausanne, Switzerland; the Intramural Research Program of the National Institute on Aging (NIA); Cancer Research United Kingdom; the UK Medical Research Council (including grants G0000649, G0000934 and G0601261); the Wellcome Trust (including Strategic Award 076113, grants 068545/Z/02 and 076467/Z/05/Z); the NIHR through the Biomedical Research Centres at Oxford, King's College London; Guys and St. Thomas' Foundation Hospitals' Trust; the British Heart Foundation (including grant FS/05/061/19501), European Community's Seventh Framework Programme (ENGAGE:HEALTH-F4-2007-201413); Diabetes UK; Unilever Corporate Research; American Diabetes Association including a Smith Family Foundation Pinnacle Program Project Award #7-03-PPG-04R; the Academy of Finland (grants 118065 and 124243); National Genome Research Net Germany; Munich Center of Health Sciences (MC Health) as part of LMUinnovativ; the Helmholtz Center Munich; the Sigrid Juselius Foundation;

University of Bristol; Linné grant from Swedish Research Council; Wallenberg Foundation; Folkhälsan Research Foundation; University of Southampton; Netherlands Organisation of Scientific Research NWO (nr. 175.010.2005.011); Erasmus Medical Center and Erasmus University, Rotterdam; Netherlands Organization for the Health Research and Development (ZonMw); the Research Institute for Diseases in the Elderly (RIDE); the Netherlands Ministry of Education, Culture and Science; the Netherlands Ministry for Health, Welfare and Sports; the European Commission (DG XII) and the Municipality of Rotterdam. G.R.A. and K.L.M. are Pew Scholars for the Biomedical Sciences; A.L.E. is supported by a Sarnoff Cardiovascular Research Foundation Fellowship; C.M.L. is a Nuffield Department of Medicine Scientific Leadership Fellow; S.A.M. is supported by a Life Sciences Research Fellowship; M.K. is supported by the Finnish Cultural Foundation; N.J.S. holds a BHF Chair; M.N.W. is a Vandervell Foundation Research Fellow; C.J.W. is supported by an American Diabetes Association postdoctoral fellowship; and E.Z. is a Wellcome Trust-RD Fellow (grant number 079557).

8.2. Supplementary Tables

rs ID	Gene	related traits	BP	p value	beta	SE	A1	A2	MAF
rs301	<i>LPL</i>	TG	19861214	0.000315	-0.14132	0.029565	G	A	0.2334
rs320	<i>LPL</i>	TG	19863357	0.000167	-0.13672	0.028204	C	A	0.2665
rs327	<i>LPL</i>	TG	19863816	0.000211	-0.13304	0.028075	C	A	0.2709
rs328	<i>LPL</i>	TG	19864004	6.60E-05	-0.23155	0.042525	G	C	0.09638
rs331	<i>LPL</i>	TG	19864685	0.000228	-0.1378	0.028386	A	G	0.2608
rs12679834	<i>LPL</i>	TG	19864713	7.30E-05	-0.22967	0.042504	G	A	0.09654
rs3208305	<i>LPL</i>	TG	19867928	0.000179	-0.13575	0.027777	T	A	0.2819
rs3735964	<i>LPL</i>	TG	19868325	3.23E-05	-0.22732	0.041032	A	C	0.1033
rs13702	<i>LPL</i>	TG	19868772	0.000244	-0.13271	0.027704	G	A	0.283
rs17482753	<i>LPL</i>	TG	19876926	8.99E-05	-0.22475	0.042387	A	C	0.09685
rs263	<i>LPL</i>	HDL-C	19857092	9.61E-05	0.127327	0.033651	A	G	0.1682
rs320	<i>LPL</i>	HDL-C	19863357	2.56E-06	0.127149	0.028236	C	A	0.2665

rs327	LPL	HDL-C	19863816	2.16E-06	0.126428	0.028103	C	A	0.2709
rs328	LPL	HDL-C	19864004	7.80E-06	0.179427	0.042637	G	C	0.09638
rs331	LPL	HDL-C	19864685	1.08E-06	0.131653	0.028414	A	G	0.2608
rs12679834	LPL	HDL-C	19864713	8.11E-06	0.179219	0.042613	G	A	0.09654
rs3208305	LPL	HDL-C	19867928	3.05E-07	0.135128	0.027797	T	A	0.2819
rs3735964	LPL	HDL-C	19868325	8.40E-06	0.174595	0.041142	A	C	0.1033
rs13702	LPL	HDL-C	19868772	2.81E-07	0.135123	0.02772	G	A	0.283
rs17482753	LPL	HDL-C	19876926	1.23E-05	0.175461	0.042493	A	C	0.09685
rs12708967	CETP	HDL-C	55550712	7.31E-08	-0.17187	0.033029	G	A	0.1834
rs3764261	CETP	HDL-C	55550825	2.41E-19	0.251125	0.027175	A	C	0.3318
rs12720918	CETP	HDL-C	55551713	2.74E-08	-0.1521	0.028181	G	A	0.2272
rs17231506	CETP	HDL-C	55552029	6.07E-19	0.249239	0.027296	G	A	0.2774
rs1800775	CETP	HDL-C	55552737	5.09E-11	0.174275	0.025822	A	C	0.4913
rs711752	CETP	HDL-C	55553712	1.97E-11	0.182519	0.026059	A	G	0.4376
rs708272	CETP	HDL-C	55553789	1.67E-11	0.182834	0.026028	A	G	0.4377
rs1864163	CETP	HDL-C	55554734	1.31E-11	-0.19809	0.029044	A	G	0.258
rs11508026	CETP	HDL-C	55556829	4.95E-11	0.179469	0.026061	A	G	0.4379
rs12720922	CETP	HDL-C	55558386	6.87E-12	-0.22576	0.033227	A	G	0.1807
rs9939224	CETP	HDL-C	55560233	6.36E-11	-0.20887	0.031698	A	C	0.2044
rs1532625	CETP	HDL-C	55562802	1.93E-12	0.18948	0.025966	A	G	0.4448
rs1532624	CETP	HDL-C	55562980	1.76E-12	0.1897	0.025959	A	C	0.4446
rs11076175	CETP	HDL-C	55563879	1.76E-11	-0.22426	0.033539	G	A	0.175
rs7499892	CETP	HDL-C	55564091	1.51E-11	-0.22431	0.0335	A	G	0.175
rs11076176	CETP	HDL-C	55564947	3.81E-07	-0.16652	3.36E-02	C	A	0.1708
rs289714	CETP	HDL-C	55564952	6.75E-06	-0.14556	0.033076	G	A	0.1784

Table 8-1A. Significant SNPs for lipids traits in BWHHS individual level data. Beta is standardized by Z score: $Z = (\text{beta} - \text{mean of beta}) / \text{SD of beta}$

rs ID	Gene	related traits	BP	p value	beta	SE	A1	A2	MAF
rs4657139	NOS1AP	QTC	160296531	1.43E-05	0.132798	0.028522	A	T	0.3382

rs10918594	NOS1AP	QTC	160297312	2.67E-06	0.148328	0.028922	C	G	0.32295
rs16847548	NOS1AP	QTC	160301898	7.57E-09	0.199317	0.033488	G	A	0.20365
rs6702936	NOS1AP	QTC	160303783	7.60E-05	0.134953	0.030263	G	A	0.28235
rs10918602	NOS1AP	QTC	160304324	5.95E-05	0.145625	0.030414	G	A	0.27585
rs10800279	NOS1AP	QTC	160304902	1.88E-05	0.142141	0.030371	G	A	0.2768
rs1415257	NOS1AP	QTC	160328668	4.42E-05	0.11301	0.028231	G	A	0.3483
rs11579850	NOS1AP	QTC	160333846	3.36E-05	0.112479	0.028182	C	A	0.34905
rs10494365	NOS1AP	QTC	160350613	2.41E-06	0.19626	0.042214	C	G	0.1154
rs10494366	NOS1AP	QTC	160352309	4.66E-05	0.112812	0.028241	C	A	0.3487
rs12733821	NOS1AP	QTC	160353315	3.33E-05	0.118197	0.028416	G	C	0.3438
rs4657154	NOS1AP	QTC	160374743	7.54E-05	0.11784	0.030417	A	G	0.26455
rs1932933	NOS1AP	QTC	160384670	6.42E-05	0.116762	0.028077	A	G	0.3524
rs945708	NOS1AP	QTC	160385526	6.35E-05	0.121033	0.02806	A	G	0.3556
rs7534004	NOS1AP	QTC	160413333	2.36E-09	0.221772	0.037865	A	G	0.15135
rs12039600	NOS1AP	QTC	160426106	3.42E-09	0.256379	0.042757	A	G	0.11575
rs4657166	NOS1AP	QTC	160427963	7.96E-06	0.12918	0.029425	C	G	0.30135
rs10918859	NOS1AP	QTC	160435892	2.89E-07	0.181159	0.035478	A	G	0.1802
rs10800352	NOS1AP	QTC	160439313	2.60E-07	0.171028	0.032977	G	A	0.21555
rs6664702	NOS1AP	QTC	160471531	3.30E-06	0.14518	0.031613	G	A	0.2444
rs10399680	NOS1AP	QTC	160480119	1.86E-06	0.151172	0.031224	G	A	0.2552
rs6427664	NOS1AP	QTC	160481097	7.15E-07	0.161283	0.031781	A	G	0.2435
rs7522678	NOS1AP	QTC	160487168	2.05E-06	0.162986	0.03369	A	C	0.201
rs10919035	NOS1AP	QTC	160510636	3.38E-05	0.16676	0.039938	A	G	0.1357
rs12053903	SCN5A	PR interval	38568397	0.000971	0.092152	0.028258	G	A	0.3177
rs10154914	SCN5A	PR interval	38607634	0.002144	-0.09657	0.033078	A	T	0.19625
rs7374540	SCN5A	PR interval	38609146	0.00347	0.077644	0.02672	C	A	0.3872
rs1805124	SCN5A	PR interval	38620424	0.008284	-0.07446	0.030758	G	A	0.2386
rs13084981	SCN5A	PR interval	38621003	0.003756	0.125451	0.042554	A	G	0.109
rs7372712	SCN5A	PR interval	38661196	0.001374	0.100148	0.033059	A	G	0.19695

rs7374540	SCN5A	QRS duration	38609146	0.00023	0.100261	0.026201	C	A	0.3849
rs9861242	SCN5A	QRS duration	38584338	0.000387	0.122806	0.029793	A	G	0.24415
rs1805126	SCN5A	QRS duration	38567410	0.000849	0.099815	0.027581	G	A	0.32195
rs10154914	SCN5A	QRS duration	38607634	0.000907	-0.11625	0.032319	A	T	0.1987
rs9833086	SCN5A	QRS duration	38585475	0.001754	0.103007	0.028261	G	A	0.279
rs7624535	SCN5A	QRS duration	38640206	0.002231	-0.09719	0.031806	C	A	0.2022
rs12053903	SCN5A	QRS duration	38568397	0.003086	0.087244	0.027786	G	A	0.31575
rs11710077	SCN5A	QRS duration	38632903	0.003849	-0.10804	0.031893	A	T	0.2097
rs12491987	SCN5A	QRS duration	38624048	0.00536	-0.10805	0.043181	A	G	0.1018

Table 8-1B. *Significant SNPs for ECG traits in BWHHS individual level data.* Effect sizes are shown as beta estimates from linear regression models for increasing copy of the coded allele and are on the standard deviation scale.

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	40433771	MFSD2	rs3103778	G	6.3	9.00E-06
2	66756976	MEIS1	rs3891585	A	2.13	1.00E-11
2	66772000	MEIS1	rs11897119	C	1.36	5.00E-11
2	215341890	VWC2L	rs7604827	C	6.54	5.00E-06
3	37574024	ITGA9	rs267567	A	2.73	4.00E-11
3	38624253	SCN5A	rs3922844	T	4.54	5.00E-43
3	38633923	SCN5A	rs11708996	C	3.04	6.00E-26
3	38681394	SCN5A	rs6763048	A	2.62	4.00E-12
3	38767315	SCN10A	rs6801957	T	3.36	9.00E-09
3	38774832	SCN10A	rs6800541	C	3.77	2.00E-74
3	65486388	MAGI1	rs1524976	A	8	8.00E-06
3	187456709	BCL6	rs3733017	G	10.07	6.00E-06
4	86641149	ARHGAP24	rs7692808	A	2.01	6.00E-20
4	86683560	ARHGAP24	rs11732231	C	2.28	3.00E-09
5	172480336	NKX2-5, C5orf41	rs251253	C	1.49	9.00E-13
6	130550063	Intergenic	rs10447419	A	8.65	2.00E-06

7	116186241	CAV1, CAV2	rs3807989	A	2.3	4.00E-28
7	116191301	CAV1	rs11773845	A	2.29	4.00E-12
10	26294828	MYO3A	rs16926523	A	9.55	4.00E-06
11	12159661	MICAL2	rs1994318	A	6.79	2.00E-06
11	75909619	WNT11	rs4944092	G	1.19	3.00E-08
12	24788339	SOX5, C12orf67	rs11047543	A	2.09	3.00E-13
12	114802138	TBX5	rs1895585	A	3.19	1.00E-19
12	115346424	TBX5, TBX3	rs1896312	C	1.95	3.00E-17
15	39573535	Intergenic	rs12595668	G	6.26	9.00E-06
15	39576560	Intergenic	rs746265	C	6.27	8.00E-06

Table 8-2. Previous GWAS findings of PR interval

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	185794883	HMCN1	rs13375391	A	5.82	9.00E-06
4	100137494	ADH6;LOC100507053	rs6820368	C	9.9	6.00E-06
5	101403586	OR7H2P - SLCO4C1	rs6861497	A	2.34	5.00E-06
6	124951063	NKAIN2	rs504008	C	2.7	3.00E-06
6	128194473	THEMIS	rs12194062	T	5.16	5.00E-06
11	67715028	RPS3AP40 - OR7E1P	rs308309	C	4.39	5.00E-06
12	69523923	CPM - CPSF6	rs10784762	T	2.38	6.00E-06
14	85737169	RNU7-51P - RNU3P3	rs1867082	A	2.82	1.00E-06
16	85961562	IRF8 - FENDRR	rs17444745	A	6.4	7.00E-06

Table 8-3. Previous GWAS findings of QRS interval

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	6279370	RNF207	rs846111	C	1.75	1.00E-16
1	162033890	OLFML2B - MIR4654	rs12143842	T	3.15	2.00E-78
1	162085685	NOS1AP	rs10494366	?	4.9	1.00E-10
1	162112910	NOS1AP	rs16857031	G	2.63	1.00E-34
1	162133117	NOS1AP	rs12029454	A	2.98	3.00E-45
1	162195738	NOS1AP	rs4657175	?	1.71	1.00E-09

1	162210610	<i>NOS1AP</i>	rs4657178	T	2.19	7.00E-33
1	169073346	<i>RPL29P7 - ATP1B1</i>	rs1320976	A	2.06	2.00E-10
1	169099483	<i>ATP1B1</i>	rs10919071	A	2.05	1.00E-15
2	40757791	<i>SLC8A1 - LDHAP3</i>	rs13017846	G	0.07	8.00E-14
2	46176962	<i>PRKCE</i>	rs6751349	A	1.22	8.00E-06
2	179248420	<i>OSBPL6</i>	rs7580640	C	3.64	3.00E-06
3	3147981	<i>IL5RA</i>	rs17879755	C	5.64	2.00E-06
3	8250790	<i>GRM7 - LMCD1-AS1</i>	rs9866825	A	1.18	8.00E-06
3	38589163	<i>SCN5A</i>	rs11129795	A	1.27	5.00E-14
3	38593393	<i>SCN5A</i>	rs12053903	C	1.23	1.00E-14
3	71115751	<i>FOXP1</i>	rs7616330	A	1.72	6.00E-06
3	73540618	<i>PDZRN3</i>	rs4557101	C	1.72	4.00E-06
4	15964863	<i>FGFBP2</i>	rs4698433	T	7.23	6.00E-06
4	16893893	<i>LDB2</i>	rs1483012	G	6.02	7.00E-06
4	16908004	<i>LDB2 - MTND5P4</i>	rs6819013	A	6.06	6.00E-06
4	27194071	<i>STIM2 - MRPL51P1</i>	rs13149020	A	10.69	8.00E-06
4	72126889	<i>SLC4A4</i>	rs2579330	C	1.36	1.00E-06
4	76030921	<i>PARM1 - RCHY1</i>	rs1426063	A	1.18	8.00E-06
4	95643957	<i>PDLIM5 - BMPR1B</i>	rs10027628	C	1.81	6.00E-06
5	3092192	<i>C5orf38 - IRX1</i>	rs6873793	C	1.47	1.00E-06
5	64417678	<i>CWC27 - RPEP1</i>	rs7729539	C	2.59	1.00E-06
5	73276903	<i>ARHGEF28 - ENC1</i>	rs6894385	C	9.92	1.00E-06
5	102759658	<i>C5orf30 - NUDT12</i>	rs17155315	G	4.73	7.00E-06
5	126032964	<i>RPLP1P7 - LMNB1</i>	rs1546498	C	7.69	2.00E-06
6	67885919	<i>NUFIP1P - BAI3</i>	rs9342616	A	6.92	1.00E-06
6	118653204	<i>SLC35F1 - BRD7P3</i>	rs12210810	C	3.13	2.00E-17
6	118680374	<i>SLC35F1 - BRD7P3</i>	rs11970286	T	1.64	2.00E-24
6	118788652	<i>CEP85L</i>	rs11752626	G	1.56	2.00E-06
6	118993632	<i>CEP85L</i>	rs11756438	A	1.4	5.00E-22

7	100881403	<i>CLDN15</i>	rs3757458	A	8	2.00E-06
7	150622162	<i>ABP1 - KCNH2</i>	rs2968864	C	1.4	8.00E-16
7	150623137	<i>ABP1 - KCNH2</i>	rs2968863	T	1.35	2.00E-15
7	150637863	<i>ABP1 - KCNH2</i>	rs4725982	T	1.58	5.00E-16
7	150669976	<i>KCNH2</i>	rs3778872	C	1.77	3.00E-06
7	154509324	<i>DPP6 - PAXIP1</i>	rs12666280	C	8.93	2.00E-06
10	30834632	<i>HNRNPA1P32 - LYZL2</i>	rs11008099	A	2.74	9.00E-06
11	2484803	<i>KCNQ1</i>	rs2074238	T	7.88	3.00E-17
11	2486120	<i>KCNQ1</i>	rs16928297	G	1.25	2.00E-06
11	2489342	<i>KCNQ1</i>	rs12296050	T	1.44	3.00E-17
11	2502319	<i>KCNQ1</i>	rs12576239	T	1.75	1.00E-15
11	2752609	<i>KCNQ1</i>	rs231906	A	1.51	2.00E-06
13	63436800	<i>RPL32P28 - OR7E156P</i>	rs2204037	A	1.38	5.00E-07
13	102553248	<i>FGF14</i>	rs9557754	G	1.23	8.00E-06
15	50873344	<i>TRPM7</i>	rs2414059	A	1.24	2.00E-06
16	8503222	<i>RPS26P51 - TMEM114</i>	rs8045405	G	11.38	9.00E-06
16	11691753	<i>LITAF - SNN</i>	rs8049607	T	1.23	5.00E-15
16	58566304	<i>CNOT1</i>	rs37060	?	1.52	1.00E-07
16	58567238	<i>CNOT1</i>	rs37062	G	1.75	3.00E-25
16	58622178	<i>CNOT1</i>	rs7188697	A	1.66	7.00E-25
17	15193056	<i>MIR4731 - TEK3</i>	rs1380181	A	1.97	3.00E-07
17	33324382	<i>LIG3</i>	rs2074518	T	1.05	6.00E-12
17	68494992	<i>CALM2P1 - SOX9</i>	rs17779747	T	1.02	6.00E-12
18	74735135	<i>MBP</i>	rs11663697	G	13.71	9.00E-06
20	47982128	<i>ZFAS1 - KCNB1</i>	rs237450	A	6.86	5.00E-06
21	37935669	<i>CLDN14</i>	rs9984896	A	2.85	5.00E-06

Table 8-4 A. Previous GWAS findings of QTc interval. QTc = QT / SQRT (RR interval)

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
-----	----------	------	----------	-------------	------	---------

1	162014632	<i>OLFML2B - MIR4654</i>	rs2880058	G	0.19	2.00E-10
1	162033890	<i>OLFML2B - MIR4654</i>	rs12143842	T	0.18	1.00E-83
2	157552860	<i>GPD2 - RPLP0P7</i>	rs7601713	A	0.16	2.00E-06
2	179641975	<i>TTN</i>	rs12476289	A	0.29	2.00E-06
3	61794054	<i>PTPRG</i>	rs652889	A	0.15	8.00E-07
3	122368347	<i>PARP15 - EIF4BP8</i>	rs2650951	A	0.29	1.00E-06
3	194327098	<i>TMEM44</i>	rs789852	A	0.25	7.00E-07
4	35410631	<i>RPL31P31 - SEC63P2</i>	rs1533317	A	0.14	2.00E-06
4	118332255	<i>TRAM1L1 - RPSAP35</i>	rs4318720	A	0.28	8.00E-07
4	148974602	<i>ARHGAP10</i>	rs6845865	G	0.19	7.00E-07
5	5897694	<i>KIAA0947 - HMGB3P3</i>	rs7728043	G	0.14	1.00E-06
6	118667522	<i>SLC35F1 - BRD7P3</i>	rs11153730	C	0.09	2.00E-29
7	37076854	<i>ELMO1</i>	rs10488031	A	0.26	2.00E-06
12	20531756	<i>PDE3A</i>	rs1348582	G	0.2	1.00E-06
13	48162558	<i>GNG5P5 - NAP1L4P3</i>	rs2478333	A	0.17	4.00E-08
14	96122408	<i>TCL6</i>	rs8015016	G	0.18	5.00E-07

Table 8-4 B. Previous GWAS findings of QT interval. Betas were calculated using unadjusted QT interval.

rs ID	Gene	related traits	BP	p value	beta	SE	Ancestral allele
rs12039600	<i>NOS1AP</i>	QTC	160426106	7.12E-23	0.256379	0.042757	G
rs7534004	<i>NOS1AP</i>	QTC	160413333	7.89E-22	0.221772	0.037865	G
rs10918859	<i>NOS1AP</i>	QTC	160435892	3.48E-20	0.181159	0.035478	G
rs10800352	<i>NOS1AP</i>	QTC	160439313	6.12E-20	0.171028	0.032977	A
rs10918594	<i>NOS1AP</i>	QTC	160297312	4.49E-18	0.148328	0.028922	C
rs7522678	<i>NOS1AP</i>	QTC	160487168	5.44E-18	0.162986	0.03369	C
rs4657139	<i>NOS1AP</i>	QTC	160296531	3.80E-17	0.132798	0.028522	A
rs6664702	<i>NOS1AP</i>	QTC	160471531	1.28E-16	0.14518	0.031613	T
rs12733821	<i>NOS1AP</i>	QTC	160353315	4.92E-16	0.118197	0.028416	C
rs10494366	<i>NOS1AP</i>	QTC	160352309	1.70E-15	0.112812	0.028241	T

rs11579850	<i>NOS1AP</i>	QTC	160333846	3.66E-15	0.112479	0.028182	G
rs1415257	<i>NOS1AP</i>	QTC	160328668	4.10E-15	0.11301	0.028231	C
rs1415262	<i>NOS1AP</i>	QTC	160312759	1.07E-14	0.094877	0.028348	C
rs5000342	<i>NOS1AP</i>	QTC	160319684	1.24E-14	0.103484	0.029562	A
rs945708	<i>NOS1AP</i>	QTC	160385526	1.50E-14	0.121033	0.02806	T
rs1932933	<i>NOS1AP</i>	QTC	160384670	1.69E-14	0.116762	0.028077	T
rs6660701	<i>NOS1AP</i>	QTC	160307530	2.28E-14	0.11512	0.028491	C
rs4657166	<i>NOS1AP</i>	QTC	160427963	1.26E-13	0.12918	0.029425	C
rs880296	<i>NOS1AP</i>	QTC	160395070	1.63E-13	0.137104	0.033702	C
rs4233385	<i>NOS1AP</i>	QTC	160314369	1.63E-13	0.0956	0.02914	C
rs10918762	<i>NOS1AP</i>	QTC	160385608	1.96E-13	0.136572	0.033563	A
rs4657154	<i>NOS1AP</i>	QTC	160374743	5.03E-12	0.11784	0.030417	N/A
rs7515045	<i>NOS1AP</i>	QTC	160310334	2.00E-11	0.097078	0.030498	C
rs6427664	<i>NOS1AP</i>	QTC	160481097	2.97E-10	0.161283	0.031781	G
rs6702936	<i>NOS1AP</i>	QTC	160303783	3.06E-10	0.134953	0.030263	A
rs4531275	<i>NOS1AP</i>	QTC	160453174	2.08E-09	0.082086	0.029558	C
rs6659759	<i>NOS1AP</i>	QTC	160304233	3.26E-09	0.133713	0.030227	T
rs12128479	<i>NOS1AP</i>	QTC	160534747	3.33E-09	0.156116	0.046353	A
rs10800366	<i>NOS1AP</i>	QTC	160470520	4.75E-09	0.075289	0.029399	C
rs10800397	<i>NOS1AP</i>	QTC	160503714	1.14E-08	0.049591	0.033375	C
rs10918602	<i>NOS1AP</i>	QTC	160304324	1.16E-08	0.145625	0.030414	T
rs10458392	<i>NOS1AP</i>	QTC	160358734	1.47E-08	0.112882	0.033224	N/A
rs10399680	<i>NOS1AP</i>	QTC	160480119	1.57E-08	0.151172	0.031224	T
rs3923368	<i>NOS1AP</i>	QTC	160463154	3.10E-08	0.061292	0.028797	T
rs10800404	<i>NOS1AP</i>	QTC	160521736	3.26E-08	0.053004	0.033504	G
rs16860185	<i>NOS1AP</i>	QTC	160560694	6.37E-08	0.150212	0.044952	G
rs7513132	<i>NOS1AP</i>	QTC	160432057	7.03E-08	0.080609	0.028532	A
rs10800279	<i>NOS1AP</i>	QTC	160304902	7.03E-08	0.142141	0.030371	T
rs10918936	<i>NOS1AP</i>	QTC	160469112	7.57E-08	0.073848	0.028029	A

rs16857031	NOS1AP	QTC	160379534	8.78E-08	0.101452	0.037934	C
rs10800409	NOS1AP	QTC	160527339	1.23E-07	0.156484	0.039781	C
rs16847548	NOS1AP	QTC	160301898	1.34E-07	0.199317	0.033488	T
rs4657161	NOS1AP	QTC	160422913	1.81E-07	0.065083	0.027944	G
rs12135795	NOS1AP	QTC	160416389	1.98E-07	0.065083	0.027944	A
rs3927640	NOS1AP	QTC	160422437	2.71E-07	0.065326	0.027953	A
rs10753765	NOS1AP	QTC	160418397	3.86E-07	0.06545	0.027937	G
rs4557949	NOS1AP	QTC	160477578	4.44E-07	0.047958	0.027479	A
rs12734991	NOS1AP	QTC	160461200	4.65E-07	0.066225	0.027712	C
rs12026452	NOS1AP	QTC	160396910	6.12E-07	0.129928	0.039262	G
rs12022557	NOS1AP	QTC	160322208	6.64E-07	0.086869	0.032799	G
rs10918615	NOS1AP	QTC	160316051	7.21E-07	0.084062	0.03284	G
rs4145621	NOS1AP	QTC	160485312	1.44E-06	0.041182	0.027544	C
rs12742393	NOS1AP	QTC	160491210	1.95E-06	0.047436	0.027862	A
rs2661818	NOS1AP	QTC	160531438	2.35E-06	0.043966	0.027617	C
rs10919035	NOS1AP	QTC	160510636	3.32E-06	0.16676	0.039938	C
rs7540690	NOS1AP	QTC	160380841	1.08E-05	0.116906	0.033688	G
rs4298709	NOS1AP	QTC	160503206	1.19E-05	0.040409	0.027957	G
rs12729882	NOS1AP	QTC	160508661	1.45E-05	0.037635	0.027877	A
rs12733377	NOS1AP	QTC	160519067	2.60E-05	0.035871	0.027888	G
rs16857019	NOS1AP	QTC	160379274	8.82E-05	0.098496	0.045361	G
rs7372712	SCN5A	PR interval	38661196	6.21E-10	0.00302	0.000488	C
rs12053903	SCN5A	PR interval	38568397	4.59E-09	0.002158	0.000368	C
rs7374540	SCN5A	PR interval	38609146	9.50E-08	0.001864	0.000349	A
rs1805126	SCN5A	PR interval	38567410	2.07E-06	0.001745	0.000368	A
rs7624535	SCN5A	PR interval	38640206	4.33E-06	-0.00194	0.000421	T
rs6768664	SCN5A	PR interval	38659470	4.75E-06	-0.00156	0.00034	A
rs7373102	SCN5A	PR interval	38655632	5.32E-06	-0.00155	0.00034	C
rs7374540	SCN5A	QRS	38609146	5.87E-09	0.100261	0.026201	A

rs9861242	SCN5A	QRS	38584338	1.59E-08	0.122806	0.029793	A
rs1805126	SCN5A	QRS	38567410	6.93E-08	0.099815	0.027581	A
rs11710077	SCN5A	QRS	38632903	1.80E-07	-0.10804	0.031893	A
rs9833086	SCN5A	QRS	38585475	2.05E-07	0.103007	0.028261	A
rs12053903	SCN5A	QRS	38568397	1.97E-06	0.087244	0.027786	C
rs6797133	SCN5A	QRS	38631037	8.73E-06	-0.05357	0.026626	G
rs7624535	SCN5A	QRS	38640206	5.17E-05	-0.09719	0.031806	T
rs9832895	SCN5A	QRS	38636537	5.60E-05	-0.02167	0.025544	C
rs7373102	SCN5A	QRS	38655632	6.45E-05	-0.05577	0.025538	C

Table 8-5. Significant SNPs for ECG traits in meta-analysis

Marker Name	Allele1	Allele2	beta	SE	P-value
rs4953023	a	g	0.7003	0.0542	3.87E-38
rs10439467	t	c	0.4272	0.0585	2.81E-13
rs4148191	a	c	0.511	0.0759	1.61E-11
rs4953019	a	g	0.3652	0.0544	1.84E-11
rs6720173	c	g	0.2636	0.0407	9.75E-11
rs2278357	t	c	0.2424	0.0406	2.30E-09
rs6756629	a	g	0.5371	0.0966	2.71E-08
rs4148189	t	c	0.2521	0.0465	5.96E-08
rs4299376	t	g	0.194	0.0359	6.44E-08
rs10208987	t	g	-0.2704	0.0527	2.90E-07
rs4077440	a	g	-0.2483	0.0526	2.38E-06
rs6709904	a	g	-0.3612	0.0779	3.58E-06
rs10201851	t	c	-0.2083	0.046	6.04E-06

Table 8-6. The significant SNPs of the GBD meta-analysis.

rsID	Locus	lipids effect size	CA effect size	direction	leading trait	CA p value	MAF	proxy	r ² with proxy
rs1084651	LPA	1.95	0.0443	same direction	HDL	0.2921	0.16	(A)	(A)
rs11869286	STARD3	-0.48	-0.0176	same direction	HDL	0.5955	0.34	(A)	(A)
rs12967135	MC4R	-0.42	0.0333	opposite direction	HDL	0.5422	0.23	rs1943226	0.188

rs1532085	<i>LIPC</i>	1.45	-0.1122	opposite direction	HDL	0.03656	0.39	(A)	(A)
rs16942887	<i>LCAT</i>	1.27	-0.0259	opposite direction	HDL	0.7018	0.12	rs2292318	0.915
rs1800961	<i>HNF4A</i>	-1.88	0.1497	opposite direction	HDL	0.0848	0.03	(A)	(A)
rs1883025	<i>ABCA1</i>	-0.94	0.0031	opposite direction	HDL	0.9312	0.25	(A)	(A)
rs2652834	<i>LACTB</i>	-0.39	0.0507	opposite direction	HDL	0.1896	0.2	rs11071721	0.442
rs2972146	<i>IRS1</i>	0.46	0.0329	same direction	HDL	0.3256	0.37	rs2943634	0.774
rs3136441	<i>LRP4</i>	0.78	-0.0428	opposite direction	HDL	0.3594	0.15	rs2070850	1
rs3764261	<i>CETP</i>	3.39	0.0259	same direction	HDL	0.444	0.32	(A)	(A)
rs386000	<i>LILRA3</i>	0.83	-0.0491	opposite direction	HDL	0.2069	0.2	rs103294	0.832
rs4129767	<i>PGS1</i>	-0.39	-0.0364	same direction	HDL	0.2487	0.49	rs4082919	0.967
rs4846914	<i>GALNT2</i>	-0.61	-0.0608	same direction	HDL	0.05724	0.4	(A)	(A)
rs6065906	<i>PLTP</i>	-0.93	0.071	opposite direction	HDL	0.07487	0.18	rs6073952	0.877
rs7134375	<i>PDE3A</i>	0.4	0.0187	same direction	HDL	0.5553	0.42	rs10841495	0.243
rs7134594	<i>MVK</i>	-0.44	0.0057	opposite direction	HDL	0.9127	0.47	(A)	(A)
rs7241918	<i>LIPG</i>	-1.31	-0.0521	same direction	HDL	0.2255	0.17	rs2156552	0.948
rs7255436	<i>ANGPTL4</i>	-0.45	-0.036	same direction	HDL	0.253	0.47	rs2278236	1
rs838880	<i>SCARB1</i>	0.61	-0.0323	opposite direction	HDL	0.352	0.31	rs838878	0.961
rs11220462	<i>ST3GAL4</i>	1.95	-0.0629	opposite direction	LDL	0.2014	0.14	rs8177375	0.121
rs1367117	<i>APOB</i>	4.05	-0.0311	opposite direction	LDL	0.358	0.3	(A)	(A)
rs1800562	<i>HFE</i>	-2.22	0.016	opposite direction	LDL	0.801	0.06	(A)	(A)
rs2479409	<i>PCSK9</i>	2.01	0.0776	same direction	LDL	0.02155	0.3	(A)	(A)
rs3757354	<i>MYLIP</i>	-1.43	-0.0129	same direction	LDL	0.8653	0.22	rs7770341	0.157
rs4299376	<i>ABCG5/8</i>	2.75	0.194	same direction	LDL	6.44E-08	0.3	(A)	(A)
rs4420638	<i>APOE</i>	7.14	0.0002	same direction	LDL	0.9977	0.17	(A)	(A)
rs6029526	<i>TOP1</i>	1.39	-0.0186	opposite direction	LDL	0.5594	0.47	rs753381	0.818
rs629301	<i>SORT1</i>	-5.65	-0.0196	same direction	LDL	0.7141	0.22	(A)	(A)
rs6511720	<i>LDLR</i>	-6.99	-0.0318	same direction	LDL	0.5239	0.11	(A)	(A)
rs7206971	<i>OSBPL7</i>	0.78	0.0457	same direction	LDL	0.1569	0.49	rs7214993	0.493
rs8017377	<i>NYNRIN</i>	1.14	0.0315	same direction	LDL	0.4356	0.47	rs6573766	0.118

rs10128711	<i>SPTY2D1</i>	-1.04	-0.0163	same direction	TC	0.6457	0.28	rs11024739	1
rs10401969	<i>CILP2</i>	-4.74	-0.0488	same direction	TC	0.6094	0.07	(A)	(A)
rs11065987	<i>BRAP</i>	-0.96	0.0334	opposite direction	TC	0.2933	0.42	rs17696736	0.696
rs1169288	<i>HNF1A</i>	1.42	0.045	same direction	TC	0.4132	0.33	(A)	(A)
rs12027135	<i>LDLRAP1</i>	-1.22	-0.0305	same direction	TC	0.5597	0.45	(A)	(A)
rs12916	<i>HMGCR</i>	2.84	0.0003	same direction	TC	0.9914	0.39	(A)	(A)
rs2000999	<i>HPR</i>	2.34	0.0878	same direction	TC	0.02439	0.2	(A)	(A)
rs2072183	<i>NPC1L1</i>	2.01	-0.0326	opposite direction	TC	0.3871	0.25	(A)	(A)
rs2081687	<i>CYP7A1</i>	1.23	0.0976	same direction	TC	0.003077	0.35	rs8192870	0.925
rs2290159	<i>RAF1</i>	-1.42	-0.0382	same direction	TC	0.3466	0.22	rs9817675	0.857
rs2902940	<i>MAFB</i>	-1.38	0.0111	opposite direction	TC	0.7313	0.29	rs6029247	0.108
rs3177928	<i>HLA</i>	2.31	-0.0547	opposite direction	TC	0.2847	0.16	rs17496549	0.646
rs492602	<i>FLJ36070</i>	1.27	-0.063	opposite direction	TC	0.2218	0.49	(A)	(A)
rs6882076	<i>TIMD4</i>	-1.98	0.0602	opposite direction	TC	0.2622	0.35	(A)	(A)
rs1042034		-5.99	-0.0172	same direction	TG	0.6511	0.22	(A)	(A)
rs11613352	<i>LRP1</i>	-2.7	-0.045	same direction	TG	0.2571	0.23	rs11172134	0.802
rs11649653	<i>CTF1</i>	-2.13	-0.0179	same direction	TG	0.5786	0.4	(A)	(A)
rs11776767	<i>PINX1</i>	2.01	-0.0367	opposite direction	TG	0.249	0.37	rs4841317	0.125
rs1260326	<i>GCKR</i>	8.76	-0.123	opposite direction	TG	0.0001345	0.41	(A)	(A)
rs12678919	<i>LPL</i>	-13.64	0.0592	opposite direction	TG	0.5069	0.12	(A)	(A)
rs1495741	<i>NAT2</i>	2.85	-0.0047	opposite direction	TG	0.893	0.22	rs1961456	0.524
rs17145738	<i>MLXIPL</i>	-9.32	0.0747	opposite direction	TG	0.1188	0.12	(A)	(A)
rs174546	<i>FADS1-2-3</i>	3.82	0.0526	same direction	TG	0.3409	0.34	(A)	(A)
rs2068888	<i>CYP26A1</i>	-2.28	-0.0309	same direction	TG	0.3302	0.46	rs4418728	1
rs2131925	<i>ANGPTL3</i>	-4.94	0.0662	opposite direction	TG	0.04468	0.32	rs1748197	1
rs2247056		-2.99	-0.004	same direction	TG	0.8998	0.25	rs2523589	0.563
rs2929282	<i>FRMD5</i>	5.13	-0.0737	opposite direction	TG	0.1577	0.05	rs10438303	0.237
rs2954029	<i>TRIB1</i>	-5.64	-0.0283	same direction	TG	0.3696	0.47	(A)	(A)
rs439401		-5.5	0.0116	opposite direction	TG	0.7225	0.36	(A)	(A)

rs442177	KLHL8	-2.25	-0.0484	same direction	TG	0.1333	0.41	rs3775214	0.964
rs5756931	PLA2G6	-1.54	0.0932	opposite direction	TG	0.004772	0.4	rs4820314	0.802
rs645040	MSL2L1	-2.22	-0.0516	same direction	TG	0.3991	0.22	(A)	(A)
rs964184	APOA1	16.95	0.0248	same direction	TG	0.6875	0.13	rs12286037	0.588

Table 8-7. *Associations between GBD and 63 independent loci previously reported to be associated with lipid levels (Teslovich et al., 2010). (A) This was not necessary, as it was directly genotyped*

2 SNPs Model						
SNP1, independent effect =1						
	N	10K	5K	1k	500	100
r ² =0.9	HR_MEAN	0.001666247	0.002327979	0.007662525	0.008899966	0.148657148
	HR_95%CI	0.03435748	0.052061965	0.118976878	0.192076277	1.452488811
	Mreg_MEAN	0.0011124	0.003888542	0.017189593	0.019894365	0.037589728
	Mreg_95%CI	0.302178051	0.438952618	0.977343389	1.423197341	3.544468234
	GCTA_MEAN	0.006184591	0.007119836	0.003234032	0.006265411	0.145259253
	GCTA_95%CI	0.04179361	0.06197972	0.146828738	0.226757081	0.832076091
r ² =0.8	HR_MEAN	0.002121781	0.002142119	0.001136835	0.005149754	0.077632799
	HR_95%CI	0.026408551	0.037983395	0.090456528	0.131271477	0.974071684
	Mreg_MEAN	0.002743894	0.002354297	0.010080058	0.008472605	0.033431403
	Mreg_95%CI	0.206969633	0.303471321	0.70384088	0.987939529	2.321674632
	GCTA_MEAN	0.002421276	0.002634457	0.000516333	0.007229987	0.061184005
	GCTA_95%CI	0.029119613	0.042309561	0.1012824	0.144272179	0.469159452
r ² =0.5	HR_MEAN	0.02039984	0.021877524	0.022225636	0.023238884	0.039157148
	HR_95%CI	0.015308343	0.02207482	0.050166436	0.07516596	0.189840069

r ² =0.2	Mreg_MEAN	0.018866975	0.022878874	0.026078252	0.018654087	0.018842517
	Mreg_95%CI	0.130271489	0.192942335	0.433039972	0.626248877	1.396102366
	GCTA_MEAN	0.020159147	0.021325037	0.021749437	0.024257734	0.044669467
	GCTA_95%CI	0.01746593	0.024185728	0.056048713	0.083885424	0.212847225
	HR_MEAN	0.000736617	0.001160796	0.002188218	0.002976302	0.01212241
	HR_95%CI	0.008099122	0.011357692	0.027778145	0.036623104	0.102899236
	Mreg_MEAN	0.000237719	0.000323408	0.00609236	0.002675935	0.006000273
	Mreg_95%CI	0.102455541	0.150891858	0.341759957	0.515294816	1.165054397
	GCTA_MEAN	0.000550163	0.000216059	0.00114004	0.001770056	0.012727879
	GCTA_95%CI	0.008818049	0.01245593	0.02935433	0.039496618	0.104033652
SNP2 independent effect =0						
r ² =0.9	N	10K	5K	1k	500	100
	HR_MEAN	0.001835105	0.002595134	0.007604252	0.008275996	0.089215875
	HR_95%CI	0.036466382	0.055232839	0.125693975	0.202003308	1.554200319
	Mreg_MEAN	0.002038094	0.004222709	0.023000853	0.020897789	0.002356261
	Mreg_95%CI	0.304112805	0.448800685	1.001041864	1.432890469	3.554742134
	GCTA_MEAN	0.006633267	0.007693014	0.002847673	0.005344696	0.146204386

r ² =0.8	GCTA_95%CI	0.044497118	0.065976995	0.155660822	0.239707121	0.859037502
	HR_MEAN	0.002449701	0.002529292	0.00067627	0.004615229	0.032919587
	HR_95%CI	0.029803049	0.042859084	0.101745378	0.146867437	1.076762496
	Mreg_MEAN	0.002580834	0.001357546	0.007358458	0.003344505	0.032221692
	Mreg_95%CI	0.209929017	0.301991991	0.712796003	0.980119523	2.362155958
	GCTA_MEAN	0.002797201	0.0031074	0.000142086	0.006755521	0.060804859
r ² =0.5	GCTA_95%CI	0.032917926	0.047824888	0.114117941	0.161942213	0.504183156
	HR_MEAN	0.013077255	0.015084665	0.014951476	0.015361469	0.029779187
	HR_95%CI	0.021624925	0.03108006	0.070312245	0.105250278	0.255825539
	Mreg_MEAN	0.01525414	0.012402253	0.024679445	0.023732623	0.038305434
	Mreg_95%CI	0.135916085	0.201657702	0.446067545	0.646961436	1.424609007
	GCTA_MEAN	0.012707242	0.014261407	0.014023917	0.016310952	0.035405846
r ² =0.2	GCTA_95%CI	0.024676355	0.034059399	0.078691106	0.117506658	0.285186833
	HR_MEAN	0.001519921	0.002343901	0.003135982	0.003716353	0.007051808
	HR_95%CI	0.018636056	0.026023183	0.06388167	0.083218802	0.217541001
	Mreg_MEAN	0.000155018	0.000107109	0.004444034	0.007841599	0.017722129
	Mreg_95%CI	0.102218767	0.151394016	0.35169023	0.510821493	1.110035301

GCTA_MEAN	0.001474759	0.00089392	0.000461927	0.000195873	0.007573781
GCTA_95%CI	0.020460327	0.028823312	0.06784384	0.090931083	0.220967026

3 SNPs model, r^2 between SNP1 and SNP2 is 0.8, r^2 between SNP1 and SNP3 are showed below

SNP1, independent effect = 1

N	SNP1 and SNP3 $r^2=0.5$						SNP1 and SNP3 $r^2=0.3$						SNP1 and SNP3 $r^2=0.1$					
	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI
10K	0.071443	0.01935	0.078972	0.213202	0.081914	0.021935	0.001721	0.020938	0.000277	0.215446	0.002296	0.024751	0.005022	0.018532	0.001694	0.208788	0.004359	0.022345
5K	0.070608	0.028433	0.074086	0.310041	0.081747	0.032022	0.001092	0.032178	0.002968	0.306323	0.00073	0.037117	0.003776	0.028718	0.001068	0.297912	0.003042	0.034279
1k	0.06793	0.068689	0.088521	0.741105	0.077826	0.080734	0.004513	0.07476	0.001779	0.666349	0.005727	0.086409	0.00339	0.063829	0.006353	0.672172	0.000589	0.075138
500	0.063203	0.101107	0.069358	1.067519	0.071937	0.111674	0.005411	0.10432	0.031336	0.989216	0.007633	0.121729	7.15E-05	0.094294	0.007536	0.992912	0.002123	0.11466
200	0.022049	0.203736	0.012563	1.573526	0.02683	0.240872	0.021471	0.185104	0.009987	1.584888	0.02954	0.224464	0.012762	0.160332	0.028184	1.598921	0.023415	0.19711
175	0.03782	0.232971	0.005792	1.77176	0.044612	0.273101	0.021517	0.193962	0.018654	1.679401	0.025672	0.228139	0.014471	0.170772	0.013852	1.682086	0.02369	0.209402
150	0.02396	0.25229	0.007221	1.914583	0.032194	0.294921	0.045519	0.92051	0.002511	1.921594	0.038327	0.262136	0.025197	0.236012	0.023125	1.906707	0.040697	0.286419
100	0.004375	0.852442	0.063161	2.372521	0.013907	0.420461	0.08198	1.688698	0.090395	2.348012	0.050037	0.357962	0.097025	0.268564	0.059514	1.952315	0.096202	0.296018

SNP2, independent effect = 0

N	SNP1 and SNP3 $r^2=0.5$						SNP1 and SNP3 $r^2=0.3$						SNP1 and SNP3 $r^2=0.1$					
	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI

10K	0.085864	0.024096	0.09158	0.260677	0.092715	0.030336	0.00245	0.026358	0.000652	0.230533	0.002878	0.031761	0.006713	0.022435	0.002014	0.212812	0.005322	0.027041
5K	0.084704	0.035927	0.083006	0.368303	0.092962	0.045254	0.001911	0.040477	0.005273	0.32278	0.001516	0.047413	0.0054	0.034904	0.002654	0.315865	0.003957	0.041322
1k	0.082365	0.086355	0.088306	0.863651	0.090444	0.11312	0.003393	0.09166	0.000757	0.733818	0.004959	0.107433	0.006199	0.077115	0.005876	0.707974	0.002631	0.09041
500	0.078792	0.122422	0.07441	1.24979	0.085522	0.154602	0.002934	0.131336	0.030136	1.094274	0.005001	0.154346	0.004201	0.11365	0.015362	1.045262	0.001601	0.13792
200	0.019562	0.236565	0.027676	1.879106	0.018457	0.303018	0.015501	0.220523	0.023563	1.75621	0.022655	0.271515	0.006207	0.188442	0.043528	1.671417	0.017751	0.231827
175	0.036375	0.267425	0.012954	2.084913	0.037855	0.346152	0.0148	0.234985	0.026581	1.852227	0.016841	0.281142	0.00682	0.203172	0.010224	1.790115	0.016837	0.24819
150	0.019694	0.287601	0.000411	2.289741	0.023012	0.367297	0.0092	0.970629	0.008168	2.065278	0.031093	0.313083	0.017137	0.267554	0.03141	1.982361	0.033987	0.324516
100	0.057824	0.920113	0.002247	2.803695	0.046178	0.493073	0.034349	1.896017	0.108352	2.51097	0.03735	0.42203	0.125322	0.306481	0.084049	2.064166	0.120576	0.345898

SNP3, independent effect = 0.3

N	SNP1 and SNP3 $r^2=0.5$						SNP1 and SNP3 $r^2=0.3$						SNP1 and SNP3 $r^2=0.1$					
	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI
10K	0.000559	0.013875	0.001203	0.167766	0.005967	0.020456	0.001627	0.017335	0.000217	0.129061	8.87E-05	0.021013	0.003715	0.017521	0.000383	0.109942	0.000112	0.019982
5K	0.000168	0.020337	0.006579	0.244004	0.005478	0.029927	0.00189	0.026149	0.001215	0.181278	0.000721	0.030983	0.003571	0.027068	0.000398	0.15904	7.9E-05	0.029661
1k	0.000423	0.048175	0.004815	0.559304	0.00439	0.07414	0.001874	0.057104	0.00599	0.440201	0.000268	0.067839	0.005076	0.05987	0.003218	0.358797	0.00135	0.066914
500	0.001545	0.065773	0.006644	0.793359	0.003692	0.103288	0.002031	0.085586	0.003048	0.613382	7.44E-05	0.103486	0.006198	0.087614	0.000279	0.523932	0.001958	0.099037
200	0.002655	0.109118	0.03065	1.28247	0.006053	0.169914	0.00746	0.134306	0.001086	0.982017	0.005463	0.162139	0.00419	0.140486	0.001762	0.82159	0.002676	0.157305
175	0.002338	0.119988	0.012632	1.394103	0.004393	0.193872	0.006821	0.144407	0.003303	1.010778	0.004826	0.177088	0.009793	0.146709	0.023451	0.891007	0.001172	0.165214
150	0.003984	0.119375	0.004589	1.529762	0.004772	0.191016	0.006389	0.15406	0.021492	1.140249	0.002143	0.184736	0.008203	0.160133	0.000956	1.003118	0.002862	0.179848

100	0.019004	0.16694	0.067913	1.793368	0.016858	0.25738	0.007147	0.315199	0.02881	1.349366	0.004199	0.232238	0.019815	0.148227	0.021706	1.095973	0.018954	0.181625
-----	----------	---------	----------	----------	----------	---------	----------	----------	---------	----------	----------	----------	----------	----------	----------	----------	----------	----------

3 SNPs model, r^2 between SNP1 and SNP2 is 0.5, r^2 between SNP1 and SNP3 are showed below

SNP1, independent effect = 1

N	SNP1 and SNP3 $r^2=0.5$						SNP1 and SNP3 $r^2=0.3$						SNP1 and SNP3 $r^2=0.1$					
	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI
10K	0.002967	0.011306	0.003693	0.14216	0.00134	0.013487	0.002058	0.010132	0.000282	0.13391	0.001713	0.012016	0.005049	0.011402	0.001047	0.136532	0.000869	0.014064
5K	0.003593	0.01675	0.000196	0.216511	0.000826	0.018758	0.002526	0.014923	0.001695	0.189744	0.001344	0.017683	0.004687	0.016598	0.005547	0.202362	0.001253	0.020857
1k	0.004979	0.03748	0.009001	0.478326	0.001466	0.042888	0.003308	0.033886	0.000837	0.42827	0.00066	0.039691	0.005785	0.037467	0.018376	0.451182	2.84E-05	0.046784
500	0.008366	0.053872	0.030902	0.680628	0.004171	0.061572	0.005471	0.05152	0.002232	0.611297	0.002761	0.057796	0.007548	0.054742	0.010261	0.619745	0.002957	0.068763
200	0.015664	0.091586	0.035961	1.111772	0.016034	0.109201	0.008398	0.080934	0.00407	0.976986	0.007091	0.098099	0.010459	0.091034	0.008501	1.005677	0.011382	0.115428
175	0.020079	0.102031	0.000426	1.175271	0.020785	0.115194	0.01516	0.088932	0.005235	1.078919	0.015724	0.10892	0.011194	0.096569	0.007069	1.102921	0.01206	0.121923
150	0.023365	0.121295	0.006161	1.217457	0.023712	0.137196	0.014484	0.096039	0.009355	1.177267	0.014081	0.115882	0.013506	0.105475	0.003759	1.146685	0.015483	0.134107

SNP2, independent effect = 0

N	SNP1 and SNP3 $r^2=0.5$						SNP1 and SNP3 $r^2=0.3$						SNP1 and SNP3 $r^2=0.1$					
	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI
10K	0.005335	0.031744	0.001524	0.165943	0.002171	0.032676	0.002866	0.020751	0.002963	0.204698	0.005029	0.031283	0.008174	0.01927	0.002284	0.162963	0.001701	0.024635

5K	0.006336	0.047868	0.002997	0.242945	0.001022	0.048058	0.003519	0.030284	0.003015	0.284677	0.004683	0.04391	0.007256	0.02827	0.001465	0.238482	0.002656	0.036918
1k	0.00701	0.100369	0.004341	0.537551	0.001246	0.104788	0.003705	0.067285	0.006813	0.627475	0.004252	0.100525	0.007309	0.063579	0.010852	0.519195	0.00314	0.08256
500	0.013831	0.148917	0.019065	0.786398	0.002959	0.15037	0.005697	0.10332	0.009217	0.936886	0.003919	0.14763	0.007902	0.091604	0.00081	0.775495	0.001821	0.119975
200	0.022976	0.243415	0.027964	1.245634	0.011116	0.25225	0.002403	0.161332	0.018789	1.500344	0.009981	0.244727	0.005543	0.150336	0.009055	1.174585	0.002733	0.198077
175	0.031973	0.261772	0.012531	1.399636	0.020975	0.274807	0.013452	0.180892	0.000335	1.563804	0.002154	0.271318	0.0045	0.157985	0.018221	1.301526	0.001541	0.20679
150	0.027984	0.310561	0.018124	1.475033	0.017212	0.304352	0.010222	0.192267	0.021124	1.701797	0.005019	0.280361	0.007534	0.172231	0.029494	1.374553	0.005823	0.226009

SNP3, independent effect = 0.3

N	SNP1 and SNP3 $r^2=0.5$						SNP1 and SNP3 $r^2=0.3$						SNP1 and SNP3 $R^2sq=0.1$					
	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI	HR_MEAN	HR_95%CI	Mreg_MEAN	Mreg_95%CI	GCTA_MEAN	GCTA_95%CI
10K	0.002248	0.024303	0.000444	0.17825	0.000404	0.026254	0.000519	0.019124	0.000176	0.178497	0.003276	0.030012	0.006345	0.017912	0.001184	0.134354	9.21E-05	0.023356
5K	0.002591	0.037651	0.005817	0.259285	0.001128	0.04049	0.000342	0.026012	0.005459	0.255805	0.003267	0.041753	0.00581	0.027432	0.002536	0.18532	0.000138	0.035817
1k	0.002004	0.079591	0.002428	0.567777	0.001139	0.091392	0.000241	0.062349	0.006501	0.572443	0.004311	0.09958	0.003768	0.061173	0.004633	0.417322	0.003319	0.076687
500	0.006028	0.118724	0.052431	0.815189	0.001748	0.125692	0.000414	0.090867	0.01073	0.855641	0.004453	0.143809	0.002437	0.085001	0.003017	0.631975	0.004808	0.112559
200	0.008509	0.186178	0.007775	1.316778	0.000498	0.208051	0.004602	0.139791	0.03704	1.347114	0.012931	0.221147	0.001143	0.140412	0.000993	0.947095	0.002521	0.179897
175	0.012829	0.20391	0.042233	1.423342	0.003982	0.232144	0.00199	0.154465	0.024355	1.380328	0.011807	0.251986	0.00372	0.15131	0.007127	0.992171	0.004315	0.195721
150	0.006318	0.238321	0.0112	1.621696	0.001548	0.255486	0.00227	0.17163	0.011046	1.476126	0.014893	0.267052	0.001402	0.158987	0.012859	1.161024	0.002415	0.203338

Table 8-8. Performance Comparison of HAPRAP, GCTA and Multiple Regression using Artificial Meta-Analyses of the Simulated Populations. **A**, errors of HAPRAP, GCTA and multiple regression in 2-SNPs models; **B**, errors of HAPRAP, GCTA and multiple regression in 3-SNPs models. “ r^2 ” is

the pair-wise LD r^2 . N is the number of individuals in the genotypes reference panel. HR_MEAN (and HR_95%CI) are the means (and width of 95% confidence intervals (CI)) for differences between HAPRAP betas and multiple regression joint SNP effects. Mreg_MEAN (and Mreg_95%CI) are the means (and width of 95% CIs) for differences between multiple regression joint SNP effects of the reference panel and multiple regression joint SNP effects of the whole population. GCTA_MEAN (and GCTA_95%CI) are the means (and width of 95% CIs) for differences between GCTA betas and multiple regression joint SNP effects.

BWHHS individual-level data														
genomic regions and traits	single SNP regression					multiple regression			HAPRAP			GCTA		
	SNP	beta	SE	pval	N	beta	SE	pval	beta	SE	pval	beta	SE	pval
APOE_TC	rs769449	0.148	0.058	1.09E-02	1962	0.097	0.058	9.60E-02	0.102	0.060	8.78E-02	0.104	0.059	7.65E-02
APOE_TC	chr19:50103919	-0.516	0.072	1.08E-12	1951	-0.504	0.072	4.23E-12	-0.502	0.071	2.05E-12	-0.501	0.073	8.26E-12
APOE_LDL	rs769449	0.174	0.053	9.81E-04	1909	0.111	0.052	3.34E-02	0.117	0.053	2.80E-02	0.120	0.053	2.44E-02
APOE_LDL	chr19:50103919	-0.632	0.065	6.35E-22	1898	-0.617	0.065	6.66E-21	-0.615	0.065	1.48E-20	-0.615	0.067	3.67E-20
NOS1AP_QTc	chr1:160291611	6.169	1.193	2.55E-07	1981	4.530	1.663	6.50E-03	4.368	1.747	1.25E-02	4.420	1.676	8.36E-03
NOS1AP_QTc	chr1:160297312	3.235	0.840	1.21E-04	1980	-2.122	1.822	2.44E-01	-2.394	1.989	2.29E-01	-1.976	1.807	2.74E-01
NOS1AP_QTc	rs12143842	5.299	0.919	9.41E-09	1981	4.262	1.928	2.71E-02	4.583	2.119	3.06E-02	4.056	1.900	3.28E-02
NOS1AP_QTc	rs16857031	3.098	1.111	5.36E-03	1981	2.815	1.591	7.70E-02	2.926	1.676	8.10E-02	2.797	1.595	7.95E-02
NOS1AP_QTc	rs10918740	3.168	0.818	1.12E-04	1981	-1.167	1.619	4.71E-01	-1.202	1.721	4.85E-01	-1.053	1.631	5.19E-01
NOS1AP_QTc	chr1:160426106	5.868	1.242	2.49E-06	1981	-0.817	2.945	7.82E-01	-0.775	3.383	8.19E-01	-0.408	2.817	8.85E-01

NOS1AP_QTc	chr1:160427963	3.544	0.855	3.55E-05	1981	0.893	1.421	5.30E-01	0.715	1.597	6.54E-01	0.922	1.419	5.16E-01
NOS1AP_QTc	chr1:160431793	4.938	1.103	8.04E-06	1981	-0.610	2.413	8.01E-01	-0.644	2.666	8.09E-01	-0.793	2.362	7.37E-01
NOS1AP_QTc	chr1:160435892	5.306	1.038	3.48E-07	1981	3.019	2.295	1.89E-01	3.135	2.662	2.39E-01	2.824	2.259	2.11E-01
NOS1AP_QTc	chr1:160477234	4.295	0.898	1.86E-06	1981	2.069	1.773	2.43E-01	2.135	1.993	2.84E-01	2.047	1.758	2.44E-01
NOS1AP_QTc	rs10919024	4.615	1.163	7.44E-05	1981	-0.738	1.948	7.05E-01	-0.861	2.137	6.87E-01	-0.574	1.940	7.67E-01
SCN5A10A_PR	chr3:38619266	-3.379	0.977	5.53E-04	1896	-2.881	0.983	3.41E-03	-2.922	0.972	2.67E-03	-2.953	0.987	2.77E-03
SCN5A10A_PR	chr3:38634956	2.607	0.867	2.68E-03	1895	2.132	0.876	1.50E-02	2.123	0.862	1.39E-02	2.110	0.881	1.66E-02
SCN5A10A_PR	chr3:38719374	2.460	0.814	2.55E-03	1895	0.707	1.040	4.96E-01	0.738	1.045	4.80E-01	0.735	1.046	4.83E-01
SCN5A10A_PR	chr3:38742319	3.283	0.825	7.13E-05	1896	2.670	1.196	2.57E-02	2.719	1.210	2.48E-02	2.627	1.199	2.84E-02
SCN5A10A_PR	chr3:38762595	-2.398	0.824	3.64E-03	1894	0.075	1.156	9.48E-01	0.158	1.173	8.93E-01	0.040	1.162	9.72E-01
LIPC_HDL	rs1077834	0.064	0.017	1.86E-04	1963	0.010	0.039	7.90E-01	0.023	0.039	5.49E-01	0.018	0.036	6.08E-01
LIPC_HDL	rs8033940	0.055	0.015	3.45E-04	1963	0.017	0.028	5.44E-01	0.010	0.029	7.38E-01	0.014	0.027	6.04E-01
LIPC_HDL	rs12914035	-0.066	0.020	1.29E-03	1955	-0.048	0.021	2.45E-02	-0.050	0.021	1.73E-02	-0.048	0.021	2.51E-02
LIPC_HDL	rs261338	0.070	0.019	1.73E-04	1964	0.073	0.044	9.66E-02	0.065	0.046	1.54E-01	0.063	0.043	1.40E-01
LIPC_HDL	rs261336	0.053	0.019	4.96E-03	1963	-0.040	0.038	2.89E-01	-0.038	0.040	3.33E-01	-0.035	0.038	3.52E-01
LPL_HDL	chr8:19852310	0.041	0.021	4.45E-02	1963	0.002	0.035	9.57E-01	0.000	0.039	9.94E-01	0.006	0.034	8.63E-01
LPL_HDL	chr8:19855067	0.046	0.027	8.27E-02	1964	0.046	0.044	2.98E-01	0.044	0.048	3.59E-01	0.034	0.043	4.28E-01
LPL_HDL	chr8:19857947	0.060	0.019	1.58E-03	1963	-0.009	0.038	8.07E-01	-0.009	0.041	8.34E-01	-0.004	0.037	9.14E-01
LPL_HDL	chr8:19861214	0.035	0.017	4.01E-02	1963	0.018	0.047	6.92E-01	0.016	0.051	7.58E-01	0.006	0.046	8.95E-01

LPL_HDL	chr8:19862322	0.027	0.016	8.48E-02	1964	-0.061	0.036	8.84E-02	-0.063	0.040	1.19E-01	-0.053	0.035	1.31E-01
LPL_HDL	chr8:19868772	0.050	0.016	1.60E-03	1963	-0.011	0.046	8.11E-01	-0.008	0.052	8.81E-01	-0.003	0.045	9.43E-01
LPL_HDL	rs1569209	0.029	0.027	2.95E-01	1962	-0.159	0.067	1.76E-02	-0.149	0.081	6.62E-02	-0.134	0.062	3.07E-02
LPL_HDL	rs10096633	0.076	0.022	7.17E-04	1964	0.206	0.059	5.30E-04	0.204	0.069	3.27E-03	0.188	0.057	9.22E-04
LPL_HDL	chr8:19914551	0.050	0.016	2.54E-03	1964	0.041	0.047	3.86E-01	0.041	0.052	4.32E-01	0.039	0.047	4.13E-01
LPL_HDL	chr8:19919284	0.051	0.014	3.78E-04	1963	0.041	0.020	4.30E-02	0.041	0.020	4.13E-02	0.040	0.020	4.81E-02
LPL_HDL	chr8:19959564	0.009	0.016	5.69E-01	1952	0.007	0.022	7.62E-01	-0.003	0.022	8.77E-01	0.004	0.022	8.52E-01
LPL_HDL	rs9644568	0.036	0.022	9.96E-02	1964	-0.013	0.037	7.34E-01	-0.003	0.040	9.49E-01	-0.006	0.036	8.65E-01
LPL_HDL	chr8:19983329	0.009	0.028	7.53E-01	1964	-0.063	0.046	1.66E-01	-0.066	0.054	2.20E-01	-0.070	0.045	1.22E-01
LDLR_TC	rs6511720	-0.059	0.060	3.31E-01	1967	0.171	0.105	1.02E-01	0.179	0.109	1.01E-01	0.170	0.104	1.03E-01
LDLR_TC	chr19:11067575	0.019	0.038	6.25E-01	1967	-0.036	0.047	4.47E-01	-0.028	0.047	5.46E-01	-0.026	0.046	5.70E-01
LDLR_TC	chr19:11068516	-0.140	0.097	1.47E-01	1968	0.021	0.125	8.69E-01	0.025	0.121	8.34E-01	0.020	0.122	8.67E-01
LDLR_TC	rs2228671	-0.139	0.059	1.79E-02	1967	-0.314	0.129	1.52E-02	-0.313	0.131	1.72E-02	-0.300	0.126	1.74E-02
LDLR_TC	rs2738447	-0.006	0.039	8.76E-01	1937	-0.033	0.052	5.20E-01	-0.030	0.051	5.57E-01	-0.027	0.051	6.00E-01
LDLR_TC	rs5742911	-0.044	0.042	2.91E-01	1967	0.016	0.056	7.73E-01	0.010	0.056	8.61E-01	0.008	0.055	8.82E-01
SMARCA4_TC	rs8099996	-0.039	0.040	3.27E-01	1968	-0.044	0.085	6.04E-01	-0.037	0.086	6.69E-01	-0.034	0.083	6.84E-01
SMARCA4_TC	rs8102273	-0.034	0.041	4.05E-01	1967	-0.050	0.118	6.70E-01	-0.063	0.121	6.01E-01	-0.067	0.116	5.62E-01
SMARCA4_TC	chr19:11044837	-0.016	0.044	7.21E-01	1968	0.095	0.095	3.20E-01	0.101	0.095	2.88E-01	0.104	0.095	2.73E-01
SMARCA4_TC	chr19:11056030	-0.036	0.049	4.58E-01	1968	0.041	0.079	6.02E-01	0.049	0.081	5.41E-01	0.047	0.078	5.49E-01

SMARCA4_TC	chr19:11059187	-0.071	0.061	2.46E-01	1967	-0.116	0.104	2.66E-01	-0.123	0.107	2.51E-01	-0.121	0.103	2.41E-01
PVRL2_TC	rs7254892	-0.333	0.115	3.80E-03	1968	-0.311	0.115	7.06E-03	-0.313	0.118	7.89E-03	-0.317	0.115	6.01E-03
PVRL2_TC	rs6857	0.124	0.052	1.62E-02	1961	0.115	0.052	2.58E-02	0.113	0.053	3.35E-02	0.115	0.052	2.59E-02
APOB_TC	chr2:21079258	-0.124	0.050	1.31E-02	1968	-0.054	0.082	5.09E-01	-0.045	0.080	5.71E-01	-0.039	0.081	6.33E-01
APOB_TC	chr2:21085700	-0.104	0.039	8.33E-03	1968	0.048	0.085	5.77E-01	0.039	0.091	6.67E-01	0.043	0.081	5.95E-01
APOB_TC	chr2:21091291	-0.118	0.077	1.24E-01	1968	0.144	0.157	3.58E-01	0.082	0.189	6.63E-01	0.085	0.150	5.71E-01
APOB_TC	rs41288783	0.743	0.464	1.09E-01	1968	0.760	0.463	1.01E-01	0.704	0.686	3.05E-01	0.766	0.467	1.01E-01
APOB_TC	rs10199768	0.126	0.040	1.84E-03	1968	-0.045	0.096	6.38E-01	-0.034	0.105	7.43E-01	-0.035	0.089	6.95E-01
APOB_TC	chr2:21099811	-0.175	0.074	1.81E-02	1968	-0.185	0.186	3.21E-01	-0.145	0.273	5.94E-01	-0.128	0.169	4.49E-01
APOB_TC	rs1367117	0.187	0.042	7.31E-06	1968	0.084	0.104	4.20E-01	0.096	0.130	4.59E-01	0.083	0.103	4.22E-01
APOB_TC	chr2:21119728	-0.160	0.172	3.52E-01	1968	-0.040	0.214	8.51E-01	0.019	0.268	9.43E-01	-0.035	0.209	8.66E-01
APOB_TC	chr2:21120715	-0.240	0.054	9.67E-06	1962	-0.144	0.147	3.27E-01	-0.121	0.224	5.91E-01	-0.155	0.133	2.43E-01
APOB_TC	rs17398765	0.249	0.070	4.15E-04	1968	0.186	0.088	3.52E-02	0.177	0.090	4.87E-02	0.186	0.088	3.54E-02
APOB_TC	rs541041	-0.198	0.051	9.80E-05	1967	0.004	0.142	9.76E-01	-0.003	0.183	9.85E-01	-0.008	0.135	9.55E-01
APOB_TC	chr2:21158700	0.169	0.042	5.45E-05	1967	0.092	0.107	3.93E-01	0.047	0.123	7.04E-01	0.083	0.106	4.29E-01
APOB_TC	chr2:21162840	-0.257	0.071	3.15E-04	1967	-0.059	0.179	7.42E-01	-0.033	0.242	8.90E-01	-0.039	0.167	8.16E-01
APOB_TC	chr2:21181140	-0.153	0.042	3.28E-04	1968	0.082	0.126	5.16E-01	0.015	0.171	9.30E-01	0.058	0.121	6.30E-01
APOB_TC	chr2:21254166	0.119	0.041	3.29E-03	1967	-0.078	0.103	4.51E-01	-0.057	0.145	6.93E-01	-0.072	0.099	4.66E-01
APOB_TC	chr2:21257334	-0.104	0.045	2.05E-02	1968	-0.026	0.113	8.18E-01	0.001	0.146	9.96E-01	-0.001	0.108	9.95E-01

APOB_TC	chr2:21284866	0.160	0.045	3.79E-04	1960	-0.076	0.121	5.30E-01	-0.067	0.152	6.61E-01	-0.089	0.115	4.42E-01
APOB_TC	chr2:21300158	-0.191	0.052	2.66E-04	1968	-0.117	0.122	3.38E-01	-0.084	0.156	5.92E-01	-0.112	0.120	3.51E-01
APOB_TC	chr2:21305109	-0.077	0.050	1.27E-01	1968	0.164	0.109	1.33E-01	0.147	0.135	2.77E-01	0.159	0.104	1.25E-01
APOB_TC	chr2:21305192	-0.185	0.050	2.25E-04	1966	-0.057	0.087	5.12E-01	-0.077	0.089	3.89E-01	-0.049	0.084	5.58E-01
APOB_TC	chr2:21305332	0.168	0.041	4.45E-05	1966	0.090	0.109	4.10E-01	0.068	0.128	5.97E-01	0.088	0.101	3.85E-01
APOB_TC	rs10198972	-0.099	0.096	3.03E-01	1968	0.063	0.138	6.48E-01	0.089	0.208	6.68E-01	0.071	0.135	5.99E-01
APOB_TC	rs312049	0.143	0.039	2.89E-04	1967	0.012	0.086	8.85E-01	0.031	0.102	7.58E-01	0.023	0.082	7.79E-01
APOB_TC	rs7571647	-0.243	0.063	1.14E-04	1968	-0.084	0.112	4.54E-01	-0.093	0.117	4.28E-01	-0.097	0.109	3.74E-01
PVRL2_LDL	rs6859	0.061	0.036	8.97E-02	1914	0.012	0.039	7.65E-01	0.012	0.038	7.51E-01	0.010	0.039	7.88E-01
PVRL2_LDL	rs283813	-0.113	0.069	9.96E-02	1911	0.066	0.090	4.61E-01	0.073	0.095	4.43E-01	0.067	0.090	4.56E-01
PVRL2_LDL	rs7254892	-0.464	0.104	9.18E-06	1915	-0.503	0.135	1.92E-04	-0.509	0.144	4.16E-04	-0.505	0.135	1.73E-04
PVRL2_LDL	rs6857	0.151	0.047	1.36E-03	1909	0.126	0.051	1.44E-02	0.123	0.051	1.72E-02	0.126	0.052	1.45E-02
TOMM40_LDL	rs157580	-0.139	0.036	1.41E-04	1905	-0.146	0.039	1.99E-04	-0.146	0.038	1.13E-04	-0.144	0.040	2.93E-04
TOMM40_LDL	rs2075650	0.147	0.050	3.02E-03	1915	0.066	0.053	2.16E-01	0.064	0.054	2.36E-01	0.072	0.053	1.79E-01
TOMM40_LDL	chr19:50092587	-0.503	0.107	2.49E-06	1915	-0.554	0.108	2.88E-07	-0.550	0.112	9.89E-07	-0.550	0.109	4.21E-07
TOMM40_LDL	chr19:50095056	0.109	0.108	3.13E-01	1914	0.051	0.109	6.37E-01	0.045	0.111	6.87E-01	0.052	0.110	6.34E-01
LDLR_LDL	rs6511720	-0.061	0.055	2.63E-01	1914	0.163	0.094	8.25E-02	0.172	0.097	7.80E-02	0.164	0.094	8.08E-02
LDLR_LDL	chr19:11068102	-0.098	0.043	2.35E-02	1914	-0.024	0.067	7.20E-01	-0.036	0.066	5.89E-01	-0.032	0.066	6.22E-01
LDLR_LDL	rs2228671	-0.142	0.053	7.89E-03	1914	-0.267	0.112	1.74E-02	-0.253	0.115	2.80E-02	-0.248	0.111	2.54E-02

LDLR_LDL	rs2738447	-0.002	0.036	9.52E-01	1885	-0.019	0.041	6.48E-01	-0.018	0.042	6.75E-01	-0.016	0.041	6.91E-01
LDLR_LDL	rs5742911	-0.037	0.038	3.23E-01	1914	0.019	0.048	6.95E-01	0.011	0.049	8.29E-01	0.011	0.048	8.25E-01
SMARCA4_LDL	rs1529729	0.028	0.035	4.21E-01	1914	0.006	0.045	8.97E-01	0.005	0.046	9.11E-01	0.006	0.045	8.89E-01
SMARCA4_LDL	rs3786725	-0.041	0.038	2.80E-01	1915	-0.052	0.084	5.36E-01	-0.051	0.084	5.40E-01	-0.053	0.085	5.29E-01
SMARCA4_LDL	chr19:11044837	-0.030	0.040	4.52E-01	1915	0.047	0.089	5.98E-01	0.045	0.087	6.03E-01	0.049	0.088	5.77E-01
SMARCA4_LDL	chr19:11056030	-0.060	0.044	1.80E-01	1915	-0.008	0.074	9.15E-01	-0.007	0.074	9.20E-01	-0.008	0.072	9.15E-01
SMARCA4_LDL	chr19:11059187	-0.079	0.055	1.56E-01	1914	-0.067	0.096	4.85E-01	-0.069	0.097	4.78E-01	-0.069	0.094	4.65E-01
CELSR2_LDL	chr1:109600244	-0.051	0.036	1.59E-01	1914	0.067	0.052	1.97E-01	0.066	0.052	2.11E-01	0.063	0.051	2.17E-01
CELSR2_LDL	rs4970834	-0.135	0.045	2.96E-03	1915	0.012	0.084	8.85E-01	0.016	0.086	8.48E-01	0.011	0.084	8.99E-01
CELSR2_LDL	chr1:109618715	-0.150	0.043	5.04E-04	1915	-0.091	0.087	2.92E-01	-0.095	0.090	2.91E-01	-0.092	0.086	2.87E-01
CELSR2_LDL	chr1:109618768	-0.120	0.038	1.65E-03	1914	-0.096	0.069	1.65E-01	-0.094	0.071	1.85E-01	-0.093	0.069	1.75E-01
CELSR2_LDL	chr1:109622407	-0.165	0.062	8.07E-03	1914	-0.065	0.083	4.34E-01	-0.070	0.081	3.87E-01	-0.069	0.083	4.08E-01
APOB_LDL	chr2:21079258	-0.148	0.046	1.22E-03	1915	-0.084	0.073	2.50E-01	-0.088	0.077	2.52E-01	-0.076	0.073	2.99E-01
APOB_LDL	chr2:21085700	-0.117	0.036	1.10E-03	1915	0.034	0.074	6.42E-01	0.045	0.077	5.59E-01	0.027	0.072	7.10E-01
APOB_LDL	chr2:21093423	0.139	0.037	1.43E-04	1915	-0.022	0.074	7.62E-01	0.003	0.076	9.67E-01	-0.019	0.072	7.88E-01
APOB_LDL	rs12713956	-0.114	0.065	7.89E-02	1915	0.088	0.129	4.97E-01	0.046	0.174	7.91E-01	0.090	0.122	4.61E-01
APOB_LDL	rs41288783	0.797	0.416	5.54E-02	1915	0.807	0.416	5.23E-02	0.715	0.493	1.47E-01	0.806	0.419	5.42E-02
APOB_LDL	chr2:21103221	-0.204	0.089	2.27E-02	1915	0.108	0.119	3.66E-01	0.100	0.134	4.56E-01	0.100	0.120	4.06E-01
APOB_LDL	rs1367117	0.194	0.038	3.34E-07	1915	0.215	0.086	1.29E-02	0.198	0.096	3.81E-02	0.199	0.085	1.91E-02

APOB_LDL	rs17398765	0.231	0.064	3.26E-04	1915	0.159	0.081	5.03E-02	0.149	0.084	7.42E-02	0.156	0.080	5.10E-02
APOB_LDL	chr2:21124828	-0.224	0.049	5.29E-06	1914	-0.138	0.125	2.70E-01	-0.099	0.180	5.85E-01	-0.138	0.118	2.42E-01
APOB_LDL	chr2:21162840	-0.243	0.065	1.77E-04	1914	0.009	0.143	9.49E-01	0.001	0.199	9.98E-01	0.009	0.137	9.46E-01
APOB_LDL	chr2:21167284	0.141	0.040	3.93E-04	1915	-0.057	0.086	5.07E-01	-0.050	0.095	5.95E-01	-0.050	0.084	5.49E-01
APOB_LDL	chr2:21181140	-0.142	0.039	2.25E-04	1915	0.102	0.090	2.56E-01	0.079	0.117	5.01E-01	0.083	0.087	3.41E-01
APOB_LDL	rs4635554	0.098	0.037	8.83E-03	1904	-0.091	0.081	2.60E-01	-0.072	0.110	5.13E-01	-0.077	0.078	3.26E-01
APOB_LDL	chr2:21257334	-0.119	0.041	3.50E-03	1915	-0.042	0.108	6.99E-01	-0.049	0.128	7.02E-01	-0.035	0.103	7.37E-01
APOB_LDL	chr2:21292834	-0.200	0.048	2.79E-05	1915	-0.189	0.109	8.24E-02	-0.163	0.135	2.28E-01	-0.174	0.107	1.05E-01
APOB_LDL	chr2:21303769	0.136	0.041	9.21E-04	1910	0.021	0.088	8.07E-01	0.009	0.119	9.40E-01	0.010	0.085	9.03E-01
APOB_LDL	chr2:21305109	-0.130	0.046	4.37E-03	1915	0.063	0.094	4.99E-01	0.066	0.112	5.52E-01	0.058	0.090	5.22E-01
APOB_LDL	chr2:21305192	-0.181	0.045	7.10E-05	1913	-0.042	0.073	5.63E-01	-0.055	0.072	4.42E-01	-0.042	0.072	5.54E-01
APOB_LDL	rs7571647	-0.228	0.057	6.79E-05	1915	-0.041	0.095	6.66E-01	-0.039	0.111	7.24E-01	-0.039	0.094	6.76E-01

Table 8-9. *Performance Comparison of HAPRAP and GCTA Using BWHHS Summary Statistics and BWHHS Individual-level Genotypes.*

BWHHS summary statistics together with 1000 Genome individual-level data														
genomic region and traits	single SNP analysis					HAPRAP			Multiple regression			GCTA		
	SNP	beta	SE	pavl	N	beta	SE	pavl	beta	SE	pavl	beta	SE	pavl
NOS1AP_QTc	rs4657166	3.544	0.855	3.55E-05	1981	1.347	1.084	2.14E-01	1.160	0.992	2.42E-01	8.923	1.030	4.54E-18
NOS1AP_QTc	rs16857031	3.098	1.111	5.36E-03	1981	2.434	1.388	7.97E-02	1.985	1.294	1.25E-01	7.895	1.272	5.38E-10
NOS1AP_QTc	rs10918859	5.306	1.038	3.48E-07	1981	3.234	1.201	7.16E-03	3.277	1.152	4.51E-03	8.062	1.179	8.14E-12
LPL_HDL	rs10096633	0.076	0.022	7.17E-04	1964	0.149	0.044	7.59E-04	0.123	0.047	8.56E-03	0.249	0.062	6.45E-05
LPL_HDL	rs35237252	0.050	0.016	2.54E-03	1964	0.065	0.047	1.65E-01	0.023	0.046	6.12E-01	0.108	0.048	2.53E-02
NOS1AP_QTc	rs10429888	6.169	1.193	2.55E-07	1981	4.297	1.739	1.35E-02	4.084	1.639	1.28E-02	6.636	1.740	1.37E-04
LPL_HDL	rs343	0.046	0.027	8.27E-02	1964	-0.001	0.046	9.77E-01	-0.001	0.040	9.90E-01	0.059	0.049	2.26E-01
TOMM40_LDL	rs157580	-0.139	0.036	1.41E-04	1905	-0.169	0.037	5.49E-06	-0.164	0.037	8.61E-06	-0.124	0.037	6.94E-04
CETP_HDL	rs5880	-0.058	0.031	6.25E-02	1964	0.036	0.034	2.80E-01	0.001	0.034	9.83E-01	0.040	0.038	2.95E-01
TOMM40_LDL	rs61679753	-0.503	0.107	2.49E-06	1915	-0.572	0.102	2.69E-08	-0.568	0.107	1.24E-07	-0.470	0.108	1.22E-05
CETP_HDL	rs11076174	-0.104	0.025	4.59E-05	1964	-0.079	0.032	1.44E-02	-0.082	0.029	4.75E-03	-0.064	0.032	4.22E-02
LDLR_TC	rs2738447	-0.006	0.039	8.76E-01	1937	-0.003	0.049	9.45E-01	-0.033	0.052	5.20E-01	-0.009	0.053	8.63E-01
APOB_LDL	rs17398765	0.231	0.064	3.26E-04	1915	0.131	0.080	1.00E-01	0.114	0.070	1.05E-01	0.143	0.069	3.89E-02
HERPUD1_HDL	rs3764261	0.103	0.015	2.29E-11	1963	0.075	0.021	4.83E-04	0.077	0.020	1.23E-04	0.084	0.020	3.66E-05
LDLR_TC	rs6511721	0.019	0.038	6.25E-01	1967	-0.014	0.046	7.54E-01	-0.036	0.047	4.47E-01	-0.021	0.047	6.56E-01
HERPUD1_HDL	rs2562126	-0.021	0.021	3.09E-01	1963	0.011	0.022	6.06E-01	0.005	0.022	8.10E-01	0.012	0.022	5.95E-01

HERPUD1_HDL	rs7205692	-0.046	0.020	2.01E-02	1964	-0.044	0.023	5.21E-02	-0.035	0.023	1.27E-01	-0.029	0.024	2.29E-01
LPL_HDL	rs301	0.035	0.017	4.01E-02	1963	0.008	0.020	7.08E-01	-0.009	0.045	8.46E-01	0.001	0.023	9.61E-01
LDLR_TC	rs6511720	-0.059	0.060	3.31E-01	1967	0.237	0.124	5.56E-02	0.171	0.105	1.02E-01	0.196	0.108	7.07E-02
HERPUD1_HDL	rs12920974	-0.077	0.016	1.38E-06	1963	-0.058	0.020	4.58E-03	-0.056	0.020	4.97E-03	-0.051	0.019	8.47E-03
APOB_TC	rs17398765	0.249	0.070	4.15E-04	1968	0.134	0.083	1.06E-01	0.142	0.077	6.73E-02	0.158	0.075	3.61E-02
PVRL2_LDL	rs7254892	-0.464	0.104	9.18E-06	1915	-0.443	0.101	1.29E-05	-0.444	0.105	2.41E-05	-0.424	0.106	6.32E-05
LDLR_LDL	rs2228671	-0.142	0.053	7.89E-03	1914	-0.137	0.066	3.73E-02	-0.149	0.060	1.33E-02	-0.140	0.057	1.37E-02
CETP_HDL	rs1864163	-0.105	0.016	1.71E-10	1964	-0.047	0.028	9.18E-02	-0.054	0.024	2.56E-02	-0.051	0.027	5.86E-02
SMARCA4_TC	rs8099996	-0.039	0.040	3.27E-01	1968	-0.033	0.083	6.87E-01	-0.044	0.085	6.04E-01	-0.032	0.087	7.11E-01
LIPC_HDL	rs261338	0.070	0.019	1.73E-04	1964	0.060	0.020	2.47E-03	0.059	0.019	1.88E-03	0.062	0.019	1.11E-03
LDLR_TC	rs73015030	-0.140	0.097	1.47E-01	1968	0.014	0.138	9.17E-01	0.021	0.125	8.69E-01	0.034	0.125	7.88E-01
SMARCA4_TC	rs73015007	-0.016	0.044	7.21E-01	1968	0.096	0.098	3.31E-01	0.095	0.095	3.20E-01	0.105	0.097	2.79E-01
PVRL2_TC	rs7254892	-0.333	0.115	3.80E-03	1968	-0.315	0.113	5.44E-03	-0.311	0.115	7.06E-03	-0.299	0.116	1.03E-02
APOB_TC	rs1367117	0.187	0.042	7.31E-06	1968	0.108	0.049	2.70E-02	0.109	0.048	2.24E-02	0.113	0.046	1.51E-02
APOE_LDL	rs769449	0.174	0.053	9.81E-04	1909	0.123	0.053	2.03E-02	0.111	0.052	3.34E-02	0.115	0.053	3.04E-02
APOE_TC	rs769449	0.148	0.058	1.09E-02	1962	0.107	0.060	7.49E-02	0.097	0.058	9.60E-02	0.100	0.059	8.77E-02
SCN5A10A_PR	rs6801957	3.283	0.825	7.13E-05	1896	3.244	0.851	1.42E-04	3.061	0.824	2.08E-04	3.106	0.832	1.90E-04
SMARCA4_TC	rs11668477	-0.036	0.049	4.58E-01	1968	0.041	0.075	5.87E-01	0.041	0.079	6.02E-01	0.045	0.075	5.43E-01
APOE_LDL	rs7412	-0.632	0.065	6.35E-22	1898	-0.614	0.069	1.57E-18	-0.617	0.065	6.66E-21	-0.614	0.067	4.45E-20

APOE_TC	rs7412	-0.516	0.072	1.08E-12	1951	-0.500	0.076	6.55E-11	-0.504	0.072	4.23E-12	-0.501	0.073	9.25E-12
LDLR_TC	rs2228671	-0.139	0.059	1.79E-02	1967	-0.343	0.134	1.07E-02	-0.314	0.129	1.52E-02	-0.310	0.126	1.39E-02
CELSR2_LDL	rs7528419	-0.150	0.043	5.04E-04	1915	-0.127	0.054	1.79E-02	-0.125	0.053	1.80E-02	-0.124	0.050	1.34E-02
SMARCA4_TC	rs17248720	-0.071	0.061	2.46E-01	1967	-0.111	0.108	3.06E-01	-0.116	0.104	2.66E-01	-0.115	0.094	2.21E-01
APOB_TC	rs1800481	-0.240	0.054	9.67E-06	1962	-0.186	0.053	4.69E-04	-0.181	0.057	1.36E-03	-0.182	0.057	1.40E-03
APOB_LDL	rs1713222	-0.224	0.049	5.29E-06	1914	-0.164	0.049	8.21E-04	-0.159	0.051	1.86E-03	-0.161	0.052	1.86E-03
SMARCA4_LDL	rs11668477	-0.060	0.044	1.80E-01	1915	-0.034	0.056	5.45E-01	-0.026	0.067	7.01E-01	-0.027	0.057	6.29E-01
SMARCA4_LDL	rs17248720	-0.079	0.055	1.56E-01	1914	-0.037	0.082	6.51E-01	-0.043	0.084	6.13E-01	-0.048	0.070	4.92E-01
SMARCA4_LDL	rs3786725	-0.041	0.038	2.80E-01	1915	-0.021	0.042	6.15E-01	-0.019	0.043	6.47E-01	-0.023	0.041	5.75E-01
CETP_HDL	rs289715	0.038	0.022	7.88E-02	1964	0.010	0.025	6.92E-01	0.014	0.022	5.32E-01	0.011	0.022	6.30E-01
SCN5A10A_PR	rs55824920	2.607	0.867	2.68E-03	1895	2.088	0.872	1.68E-02	2.171	0.868	1.25E-02	2.048	0.877	1.95E-02
LIPC_HDL	rs12914035	-0.066	0.020	1.29E-03	1955	-0.054	0.020	7.87E-03	-0.051	0.021	1.37E-02	-0.054	0.021	8.60E-03
APOB_LDL	rs1367117	0.194	0.038	3.34E-07	1915	0.103	0.046	2.41E-02	0.126	0.043	3.65E-03	0.119	0.043	5.82E-03
CELSR2_LDL	rs11577931	-0.165	0.062	8.07E-03	1914	-0.056	0.080	4.84E-01	-0.060	0.076	4.29E-01	-0.074	0.072	3.03E-01
SMARCA4_TC	rs8102273	-0.034	0.041	4.05E-01	1967	-0.066	0.115	5.64E-01	-0.050	0.118	6.70E-01	-0.074	0.119	5.35E-01
PVRL2_TC	rs6857	0.124	0.052	1.62E-02	1961	0.112	0.054	3.67E-02	0.115	0.052	2.58E-02	0.104	0.052	4.57E-02
SCN5A10A_PR	rs62242769	-3.379	0.977	5.53E-04	1896	-3.450	1.064	1.21E-03	-2.931	0.979	2.79E-03	-3.197	0.983	1.15E-03
PVRL2_LDL	rs6857	0.151	0.047	1.36E-03	1909	0.134	0.049	6.08E-03	0.138	0.047	3.17E-03	0.123	0.048	9.94E-03
CETP_HDL	rs118146573	-0.098	0.022	5.60E-06	1964	-0.075	0.036	3.54E-02	-0.056	0.032	8.53E-02	-0.067	0.035	5.20E-02

LDLR_LDL	rs5742911	-0.037	0.038	3.23E-01	1914	-0.010	0.042	8.07E-01	0.012	0.043	7.87E-01	-0.004	0.040	9.29E-01
HERPUD1_HDL	rs9938160	0.022	0.015	1.60E-01	1964	-0.019	0.017	2.79E-01	-0.012	0.016	4.67E-01	-0.019	0.017	2.65E-01
LPL_HDL	rs74304285	0.041	0.021	4.45E-02	1963	-0.028	0.043	5.13E-01	-0.003	0.034	9.24E-01	-0.024	0.042	5.70E-01
LPL_HDL	rs4333617	0.051	0.014	3.78E-04	1963	0.045	0.020	2.15E-02	0.040	0.019	4.17E-02	0.029	0.019	1.25E-01
LDLR_TC	rs5742911	-0.044	0.042	2.91E-01	1967	-0.023	0.051	6.55E-01	0.016	0.056	7.73E-01	-0.016	0.052	7.52E-01
CETP_HDL	rs5883	0.118	0.031	1.64E-04	1962	0.089	0.029	2.39E-03	0.090	0.032	4.26E-03	0.070	0.033	3.32E-02
APOB_LDL	rs41288783	0.797	0.416	5.54E-02	1915	0.705	0.206	6.39E-04	0.731	0.413	7.70E-02	0.449	0.429	2.96E-01
LPL_HDL	rs9644568	0.036	0.022	9.96E-02	1964	-0.010	0.036	7.76E-01	-0.012	0.035	7.19E-01	-0.040	0.044	3.64E-01
LPL_HDL	rs269	0.060	0.019	1.58E-03	1963	0.024	0.045	5.94E-01	0.023	0.035	5.07E-01	-0.006	0.042	8.89E-01
NOS1AP_QTc	rs12143842	5.299	0.919	9.41E-09	1981	1.610	1.441	2.64E-01	1.586	1.332	2.34E-01	0.540	1.369	6.93E-01
LPL_HDL	rs13702	0.050	0.016	1.60E-03	1963	-0.013	0.050	7.91E-01	0.013	0.044	7.73E-01	-0.026	0.046	5.66E-01
LPL_HDL	rs314	0.027	0.016	8.48E-02	1964	-0.079	0.032	1.30E-02	-0.041	0.034	2.39E-01	-0.087	0.032	6.74E-03
LPL_HDL	rs7015766	0.009	0.028	7.53E-01	1964	-0.113	0.043	8.72E-03	-0.106	0.042	1.17E-02	-0.232	0.061	1.28E-04

Table 8-10. Performance Comparison of HAPRAP and GCTA Using BWHHS Summary Statistics and 1000 Genome Individual-Level Genotypes.

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	40028180	<i>MACF1,PABPC4</i>	rs4660293	G	0.48	4.00E-10
1	182168885	<i>ZNF648</i>	rs1689800	G	0.47	3.00E-10
1	230294916	<i>GALNT2</i>	rs2144300	T	1.11	3.00E-14
1	230295691	<i>GALNT2</i>	rs4846914	G	0.61	4.00E-21
1	230304988	<i>GALNT2</i>	rs10489615	G	0.02	4.00E-09
2	21198900	<i>APOB</i>	rs11902417	G	0.02	4.00E-07
2	21206183	<i>APOB</i>	rs6754295	C	0.07	4.00E-08
2	21225281	<i>APOB</i>	rs1042034	C	0.9	1.00E-30
2	165540800	<i>COBLL1</i>	rs12328675	C	0.68	3.00E-10
2	227100698	<i>IRS1</i>	rs2972146	C	0.46	2.00E-09
4	103188709	<i>SLC39A8</i>	rs13107325	T	0.84	7.00E-11
5	53298025	<i>ARL15</i>	rs6450176	A	0.49	5.00E-08
6	34552797	<i>C6orf106</i>	rs2814944	A	0.49	4.00E-09
6	139829666	<i>CITED2</i>	rs605066	C	0.39	3.00E-08
6	161089817	<i>LPA</i>	rs1084651	A	0.56	3.00E-08
7	72982874	<i>MLXIPL</i>	rs17145738	T	0.57	1.00E-09
7	80340622	<i>CD36</i>	rs2366858	C	0.03	6.00E-10
7	130433384	<i>KLF14</i>	rs4731702	T	0.59	1.00E-15
8	9183358	<i>PPP1R3B</i>	rs9987289	A	1.21	6.00E-25
8	9185146	<i>PPP1R3B</i>	rs2126259	A	0.11	1.00E-06
8	9187242	<i>PPP1R3B</i>	rs1461729	T	0.04	7.00E-09
8	19819328	<i>LPL</i>	rs325	T	0.05	8.00E-26
8	19819439	<i>LPL</i>	rs326	T	0.02	1.00E-08
8	19819724	<i>LPL</i>	rs328	G	0.17	9.00E-23
8	19832646	<i>LPL</i>	rs17482753	T	2.02	3.00E-11
8	19844222	<i>LPL</i>	rs12678919	G	2.25	1.00E-97
8	19847690	<i>LPL</i>	rs10503669	T	0.04	8.00E-43
8	19865175	<i>LPL</i>	rs2083637	G	0.11	6.00E-18

8	116599199	<i>TRPS1</i>	rs2293889	T	0.44	6.00E-11
8	126490972	<i>TRIB1</i>	rs2954029	T	0.61	6.00E-19
9	15289578	<i>TTC39B</i>	rs471364	C	0.08	3.00E-10
9	15296034	<i>TTC39B</i>	rs643531	C	0.01	7.00E-09
9	15305378	<i>TTC39B</i>	rs581080	C	0.72	1.00E-13
9	107620835	<i>NIPSNAP3A,NIPSNAP3B,ABCA1</i>	rs9282541	A	0.37	6.00E-26
9	107647220	<i>ABCA1</i>	rs4149268	C	0.82	1.00E-10
9	107647655	<i>ABCA1</i>	rs3890182	A	0.1	3.00E-10
9	107653426	<i>ABCA1</i>	rs12686004	T	0.02	2.00E-18
9	107657070	<i>ABCA1</i>	rs3905000	G	0.11	9.00E-13
9	107664301	<i>ABCA1</i>	rs1883025	T	0.94	2.00E-33
11	10388782	<i>ADM,AMPD3</i>	rs2923084	G	0.41	5.00E-08
11	46743247	<i>LRP4,NR1H3</i>	rs3136441	C	0.78	3.00E-18
11	47286290	<i>NR1H3</i>	rs7120118	G	0.04	4.00E-08
11	48518893	<i>MADD,FOLH1</i>	rs7395662	G	0.07	6.00E-11
11	61569830	<i>FADS1,FADS2,FADS3</i>	rs174546	T	0.73	2.00E-22
11	61570783	<i>FADS1,FADS2,FADS3</i>	rs174547	C	0.09	2.00E-12
11	61571348	<i>FADS1</i>	rs174548	G	0.01	1.00E-12
11	61597212	<i>FADS2,FADS3</i>	rs174570	G	0.06	4.00E-06
11	116617240	<i>Intergenic</i>	rs11216126	C	0.03	3.00E-34
11	116648917	<i>APOA1,APOC3,APOA4,APOA5</i>	rs964184	G	1.5	5.00E-47
11	122522375	<i>UBASH3B</i>	rs7941030	G	0.31	3.00E-08
12	20473758	<i>PDE3A</i>	rs7134375	A	0.4	4.00E-08
12	57792580	<i>LRP1</i>	rs11613352	T	0.46	2.00E-08
12	109840940	<i>MYO1H,KCTD10,UBE3B,MMAB,MVK</i>	rs9943753	G	0.02	3.00E-06
12	109895168	<i>MMAB,MVK</i>	rs2338104	C	0.07	1.00E-10
12	110000193	<i>MMAB,MVK</i>	rs7134594	C	0.44	7.00E-15
12	111414461	<i>MYL2</i>	rs12229654	G	0.03	3.00E-23
12	112645401	<i>C12orf51</i>	rs2074356	T	0.04	7.00E-37

12	113409176	<i>OAS3</i>	rs2072134	A	0.02	6.00E-06
12	123796238	<i>SBNO1</i>	rs4759375	T	0.86	8.00E-09
12	124460167	<i>CCDC92,ZNF664</i>	rs4765127	T	0.44	3.00E-10
12	125261593	<i>SCARB1</i>	rs838880	C	0.61	3.00E-14
15	58674695	<i>LIPC</i>	rs4775041	C	1.38	3.00E-20
15	58678512	<i>LIPC</i>	rs10468017	T	0.1	8.00E-23
15	58683366	<i>LIPC</i>	rs1532085	A	1.45	3.00E-96
15	58694020	<i>Intergenic</i>	rs16940212	T	0.02	1.00E-24
15	58723426	<i>LIPC</i>	rs1077835	A	0.17	2.00E-14
15	58723675	<i>LIPC</i>	rs1800588	T	0.14	2.00E-32
15	58726744	<i>LIPC</i>	rs261334	G	0.03	5.00E-22
15	63396867	<i>LACTB</i>	rs2652834	A	0.39	9.00E-09
16	56985139	<i>CETP</i>	rs9989419	G	0.04	1.00E-32
16	56988044	<i>CETP</i>	rs173539	T	0.25	4.00E-75
16	56990716	<i>CETP</i>	rs247617	A	0.06	1.00E-44
16	56993324	<i>CETP</i>	rs3764261	A	3.39	7E-380
16	56995236	<i>CETP, NUP93, SLC12A3, HERPUD1</i>	rs1800775	A	3.09	4.00E-93
16	56997233	<i>CETP</i>	rs1864163	G	4.12	7.00E-39
16	57005479	<i>CETP</i>	rs1532624	C	0.21	9.00E-94
16	57012379	<i>CETP</i>	rs12708980	C	0.04	2.00E-28
16	67708897	<i>GFOD2,LCAT</i>	rs12449157	G	0.02	2.00E-07
16	67902070	<i>CTCF, PRMT8</i>	rs2271293	G	0.13	8.00E-16
16	67928042	<i>LCAT</i>	rs16942887	A	1.27	8.00E-33
16	68013471	<i>LCAT</i>	rs255049	G	0.05	3.00E-08
16	68024995	<i>LCAT</i>	rs255052	A	0.74	1.00E-07
16	81534790	<i>CMIP</i>	rs2925979	T	0.45	2.00E-11
17	2428508	<i>Intergenic</i>	rs9891572	A	0.05	2.00E-07
17	37813856	<i>STARD3</i>	rs11869286	C	0.51	3.00E-14
17	66875294	<i>ABCA8</i>	rs4148008	G	0.42	2.00E-10

17	76403984	<i>PGS1</i>	rs4129767	G	0.4	5.00E-09
18	47159090	<i>LIPG</i>	rs7240405	A	2.27	5.00E-10
18	47160953	<i>LIPG</i>	rs7241918	G	1.31	3.00E-49
18	47167214	<i>LIPG</i>	rs4939883	T	0.14	7.00E-15
18	47181668	<i>LIPG</i>	rs2156552	T	0.03	2.00E-12
18	57849023	<i>RPS3A,MC4R</i>	rs12967135	A	0.42	7.00E-09
19	8433196	<i>ANGPTL4</i>	rs7255436	C	0.45	3.00E-08
19	8469738	<i>ANGPTL4</i>	rs2967605	T	0.12	1.00E-08
19	11342703	<i>LOC55908</i>	rs12979813	T	0.02	2.00E-09
19	11347493	<i>DOCK6,LOC55908</i>	rs737337	C	0.64	3.00E-09
		<i>TSPAN16,SPC24,RAB3D,LOC55908,KANK</i>				
19	11350488	<i>2,DOCK6</i>	rs2278426	?	0.14	3.00E-09
19	45376284	<i>PVRL2, TOMM40, APOE</i>	rs519113	C	0.02	8.00E-11
19	45395266	<i>TOMM40, APOE</i>	rs157580	G	0.06	4.00E-07
19	45422946	<i>APOE,APOC1,APOC2</i>	rs4420638	G	1.06	4.00E-21
19	54792761	<i>LILRA3,LILRB2</i>	rs386000	C	0.83	4.00E-16
20	43042364	<i>HNF4A</i>	rs1800961	T	1.88	1.00E-15
20	44554015	<i>PLTP</i>	rs6065906	C	0.93	2.00E-22
20	44576502	<i>PLTP</i>	rs7679	C	0.07	4.00E-09
21	32060490	<i>Intergenic</i>	rs13046373	C	0.02	4.00E-06
22	21932068	<i>UBE2L3</i>	rs181362	T	0.46	1.00E-08

Table 8-11. Genetic variants reported to be associated with HDL-C by previous GWASs

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	25775733	<i>TMEM57,LDLRAP1</i>	rs12027135	A	1.1	1.00E-10
1	55496039	<i>PCSK9</i>	rs11206510	T	3.04	4.00E-11
1	55504650	<i>PCSK9</i>	rs2479409	G	2.01	2.00E-28
1	55505647	<i>PCSK9</i>	rs11591147	T	0.47	2.00E-44
1	55625548	<i>PCSK9</i>	rs17111684	A	9.01	2.00E-17
1	63025942	<i>ANGPTL3,DOCK7</i>	rs2131925	G	1.59	3.00E-18

1	63118196	<i>DOCK7</i>	rs10889353	C	0.06	8.00E-06
1	109817590	<i>CELSR2, PSRC1, SORT1</i>	rs12740374	T	0.23	2.00E-42
1	109817838	<i>CELSR2</i>	rs660240	A	0.04	1.00E-26
1	109818306	<i>CELSR2, PSRC1, SORT1</i>	rs629301	G	5.65	1.00E-170
1	109818530	<i>CELSR2, PSRC1, SORT1</i>	rs646776	C	0.16	3.00E-29
1	109822166	<i>CELSR2, PSRC1</i>	rs599839	G	0.16	1.00E-33
1	207875175	<i>CR1L</i>	rs4844614	A	0.1	2.00E-07
1	220973563	<i>MOSC1</i>	rs2642442	T	1.09	6.00E-11
1	234858597	<i>IRF2BP2, TOMM20</i>	rs514230	A	1.13	9.00E-12
2	21232195	<i>APOB</i>	rs693	A	0.12	1.00E-21
2	21241505	<i>APOB</i>	rs12713956	G	4.86	4.00E-08
2	21263900	<i>APOB</i>	rs1367117	A	4.05	4.00E-114
2	21286057	<i>APOB</i>	rs515135	T	0.16	5.00E-29
2	21288321	<i>APOB</i>	rs562338	G	4.89	6.00E-22
2	44065090	<i>ABCG5</i>	rs6756629	G	0.16	3.00E-10
2	44072576	<i>ABCG5, ABCG8</i>	rs4299376	G	2.75	2.00E-47
2	44073881	<i>ABCG8</i>	rs6544713	T	0.15	2.00E-20
2	44074431	<i>ABCG8</i>	rs4245791	G	5.97	1.00E-09
5	74625487	<i>HMGCR</i>	rs7703051	A	18	1.00E-08
5	74648603	<i>HMGCR</i>	rs12654264	T	0.1	1.00E-20
5	74651084	<i>HMGCR</i>	rs3846662	G	0.08	2.00E-11
5	74655726	<i>HMGCR</i>	rs3846663	T	0.07	8.00E-12
5	74656539	<i>HMGCR</i>	rs12916	C	2.45	5.00E-45
5	156390297	<i>TIMD4, HAVCR1</i>	rs6882076	T	1.67	2.00E-22
5	156398169	<i>TIMD4, HAVCR1</i>	rs1501908	G	0.07	1.00E-11
6	16127407	<i>IDOL</i>	rs3757354	T	1.43	1.00E-11
6	16197194	<i>MYLIP, GMPR</i>	rs2142672	C	0.01	2.00E-08
6	26093141	<i>HFE, HIST1H4C</i>	rs1800562	A	2.22	6.00E-10
6	32412435	<i>HLA</i>	rs3177928	A	1.83	2.00E-15

6	33143948	<i>B3GALT4</i>	rs2254287	G	1.91	5.00E-08
6	116312893	<i>FRK</i>	rs9488822	T	0.89	3.00E-09
6	160578860	<i>LPA</i>	rs1564348	C	1.95	2.00E-17
7	21607352	<i>DNAH11</i>	rs12670798	C	1.26	7.00E-10
7	44579180	<i>NPC1L1</i>	rs2072183	A	1.17	4.00E-11
8	9183358	<i>PPP1R3B</i>	rs9987289	T	2.22	7.00E-15
8	59388565	<i>CYP7A1</i>	rs2081687	A	0.95	4.00E-09
8	145043543	<i>PLEC1</i>	rs11136341	G	1.4	4.00E-13
10	113933886	<i>GPAM</i>	rs2255141	A	1.08	2.00E-09
11	61569830	<i>FADS1,FADS2,FADS3</i>	rs174546	T	1.71	1.00E-21
11	61597212	<i>FADS2, FADS3</i>	rs174570	G	0.11	4.00E-13
11	116603724	<i>APOA1, APOA4, APOA5, APOC3</i>	rs12272004	C	0.18	5.00E-13
11	116607437	<i>BUD13, ZNF259, APOA5,APOA4,APOC3,APOA1</i>	rs1558861	T	0.03	2.00E-06
11	116648917	<i>APOA1,APOC3,APOA4,APOA5</i>	rs964184	G	2.85	1.00E-26
11	126243952	<i>ST3GAL4</i>	rs11220462	A	1.95	1.00E-15
12	112072424	<i>BRAP</i>	rs11065987	G	0.97	2.00E-09
12	121388962	<i>HNF1A</i>	rs2650000	A	0.07	2.00E-08
12	121416650	<i>HNF1A</i>	rs1169288	C	1.42	1.00E-15
14	24883887	<i>CBLN3,KIAA0323</i>	rs8017377	T	1.17	4.00E-11
16	56993324	<i>CETP</i>	rs3764261	T	1.45	9.00E-13
16	72108093	<i>HP,HPR,DHX38</i>	rs2000999	A	2	2.00E-22
17	45425115	<i>OSBPL7</i>	rs7206971	T	0.87	4.00E-09
19	11195030	<i>LDLR</i>	rs11668477	G	0.13	2.00E-07
19	11200008	<i>LDLR</i>	rs17249141	T	32.93	2.00E-17
19	11202306	<i>LDLR</i>	rs6511720	T	6.99	4.00E-117
19	11210912	<i>LDLR</i>	rs2228671	G	0.14	4.00E-14
19	11238473	<i>LDLR</i>	rs2738459	C	0.02	7.00E-06
19	19407718	<i>CSPG3,CILP2,PBX4</i>	rs10401969	C	3.11	7.00E-22
19	19658472	<i>NCAN,CILP2</i>	rs16996148	G	3.32	3.00E-09

19	19789528	NCAN	rs2304130	G	0.12	3.00E-06
19	45395266	TOMM40, APOE	rs157580	G	0.11	2.00E-19
19	45403412	APOE	rs1160985	C	6.77	2.00E-21
19	45412079	APOE	rs7412	?	12.3	2.00E-09
19	45422946	APOE,APOC1,APOC2	rs4420638	G	7.14	9.00E-147
20	39091487	MAFB	rs2902940	G	0.98	1.00E-08
20	39228784	MAFB	rs6102059	T	0.06	4.00E-09
20	39672618	TOP1	rs6029526	T	1.41	3.00E-19
23	66942625	AR	rs5031002	A	0.3	2.00E-07

Table 8-12. Genetic variants reported to be associated with LDL-C by previous GWASs

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	62931632	DOCK7	rs1167998	C	0.09	2.00E-12
1	62996838	DOCK7,ANGPTL3	rs1168013	G	0.04	6.00E-08
1	63025942	ANGPTL3,DOCK7	rs2131925	G	4.94	9.00E-43
1	63049593	ANGPTL3	rs1748195	C	7.12	2.00E-10
1	63118196	ANGPTL3	rs10889353	C	0.13	2.00E-09
1	63191777	ANGPTL3, DOCK7, ATG4C	rs12130333	T	0.11	2.00E-08
1	230294916	GALNT2	rs2144300	C	4.25	8.00E-07
1	230295691	GALNT2	rs4846914	G	0.08	7.00E-15
2	21204025	APOB	rs6544366	T	0.04	2.00E-07
2	21206183	APOB	rs6754295	C	0.08	3.00E-08
2	21208211	APOB	rs7557067	G	0.08	9.00E-12
2	21225281	APOB	rs1042034	C	5.99	1.00E-45
2	21232195	APOB	rs693	A	0.08	2.00E-07
2	27598097	GCKR	rs4665972	T	0.07	1.00E-08
2	27730940	GCKR	rs1260326	T	8.76	6.00E-133
2	27741237	GCKR	rs780094	T	8.59	6.00E-32
2	27743154	GCKR	rs780092	G	0.05	5.00E-27

2	27748624	<i>GCKR</i>	rs1260333	C	0.05	2.00E-19
2	165513091	<i>COBLL1</i>	rs10195252	C	2.01	2.00E-10
2	227100698	<i>IRS1</i>	rs2972146	C	1.89	2.00E-08
3	135926622	<i>MSL2L1</i>	rs645040	G	2.22	3.00E-08
4	88030261	<i>AFF1, KLHL8</i>	rs442177	G	2.25	9.00E-12
5	55854153	<i>C5orf35</i>	rs6867983	T	0.02	3.00E-06
5	55861786	<i>ANKRD55, MAP3K1</i>	rs9686661	T	2.57	1.00E-10
5	156390297	<i>TIMD4, HAVCR1</i>	rs6882076	G	2.63	4.00E-12
6	31265490	<i>HLA</i>	rs2247056	T	2.99	2.00E-15
7	72129667	<i>TYW1B</i>	rs13238203	T	7.91	1.00E-09
7	72856269	<i>MLXIPL</i>	rs2240466	G	0.14	1.00E-12
7	72856430	<i>BAZ1B, BCL7B, TBL2, MLXIPL</i>	rs1178979	A	0.05	2.00E-12
7	72864869	<i>MLXIPL</i>	rs714052	G	0.16	3.00E-15
7	72982874	<i>MLXIPL</i>	rs17145738	G	7.91	9.00E-59
7	72987354	<i>TBL2, MLXIPL</i>	rs2286276	T	0.07	1.00E-15
7	73020337	<i>MLXIPL</i>	rs3812316	C	10.5	1.00E-10
8	10683929	<i>PINX1, XKR6</i>	rs11776767	C	2.01	1.00E-08
8	11045161	<i>XKR6, AMAC1L2</i>	rs7819412	G	0.04	3.00E-08
8	18272881	<i>NAT2</i>	rs1495741	G	2.97	4.00E-14
8	19819724	<i>LPL</i>	rs328	G	0.19	2.00E-28
8	19827848	<i>LPL</i>	rs10105606	C	0.07	4.00E-26
8	19830921	<i>LPL</i>	rs10096633	G	0.17	2.00E-18
8	19844222	<i>LPL</i>	rs12678919	G	13.64	2.00E-115
8	19847690	<i>LPL</i>	rs10503669	T	0.09	7.00E-39
8	19848080	<i>LPL</i>	rs17410962	G	0.11	7.00E-09
8	126477978	<i>Intergenic</i>	rs2001945	C	0.04	1.00E-20
8	126486409	<i>TRIB1</i>	rs17321515	G	0.08	4.00E-17
8	126490972	<i>TRIB1</i>	rs2954029	T	5.64	3.00E-55
10	65027610	<i>JMJD1C</i>	rs10761731	T	2.38	3.00E-12

10	94839642	<i>CYP26A1</i>	rs2068888	A	2.28	2.00E-08
11	61569830	<i>FADS1,FADS2,FADS3</i>	rs174546	T	3.82	5.00E-24
11	61570783	<i>FADS1, FADS2, FADS3</i>	rs174547	C	0.06	2.00E-14
11	61571348	<i>FADS1</i>	rs174548	G	0.03	5.00E-14
11	116584987	<i>BUD13,ZNF259,APOA5,APOA4,APOC3,APOA1</i>	rs4938303	T	0.07	4.00E-21
11	116586283	<i>APOA1,APOC3,APOA4,APOA5</i>	rs7350481	G	0.24	1.00E-49
11	116603724	<i>APOA1, APOA4, APOA5, APOC3</i>	rs12272004	C	0.18	5.00E-13
11	116607437	<i>LOC440069, MGC13125</i>	rs1558861	C	17	2.00E-26
11	116619073	<i>APOA1, APOC3, APOA4, APOA5, ZNF259, BUD13</i>	rs28927680	G	0.26	2.00E-17
11	116648917	<i>APOA1,APOC3,APOA4,APOA5</i>	rs964184	G	16.95	7.00E-240
11	116652207	<i>APOA5, APOA4, APOC3, APOA1</i>	rs12286037	T	25.82	1.00E-26
11	116652423	<i>APOA,APOC</i>	rs6589566	C	0.21	5.00E-14
11	116654435	<i>ZNF259, APOA1, APOC3, APOA4, APOA5, BUD13</i>	rs603446	T	0.09	2.00E-86
11	116732512	<i>APOA1,KIAA0999,LOC645044</i>	rs2075292	G	8.7	5.00E-08
12	57792580	<i>LRP1</i>	rs11613352	T	2.7	4.00E-10
12	124460167	<i>CCDC92,ZNF664</i>	rs4765127	G	2.42	1.00E-08
15	42683787	<i>CAPN3</i>	rs2412710	A	7	2.00E-08
15	44245931	<i>FRMD5</i>	rs2929282	T	5.13	2.00E-11
15	58674695	<i>LIPC</i>	rs4775041	C	3.62	2.00E-08
15	58683366	<i>LIPC</i>	rs1532085	G	2.99	2.00E-13
16	30918487	<i>CTF1</i>	rs11649653	G	2.13	3.00E-08
16	56993324	<i>CETP</i>	rs3764261	A	2.88	1.00E-12
16	56995236	<i>CETP</i>	rs1800775	C	2.1	3.00E-13
19	19407718	<i>CSPG3,CILP2,PBX4</i>	rs10401969	C	7.83	2.00E-29
19	19658472	<i>NCAN,CILP2</i>	rs16996148	G	6.1	3.00E-09
19	19662220	<i>NCAN,CILP2,PBX4</i>	rs17216525	T	0.11	4.00E-11
19	19789528	<i>CILP2,ZNF101</i>	rs2304130	G	0.07	4.00E-08
19	45414451	<i>APOE,APOC1,APOC2</i>	rs439401	T	5.5	1.00E-30
19	45422587	<i>APOC1</i>	rs12721054	G	0.1	3.00E-19

19	45422946	<i>APOE cluster</i>	rs4420638	G	2.4	3.00E-13
20	44554015	<i>PLTP</i>	rs6065906	C	3.32	5.00E-18
20	44576502	<i>PLTP</i>	rs7679	C	0.07	7.00E-11
22	38546033	<i>PLA2G6</i>	rs5756931	C	1.54	4.00E-08

Table 8-13. Genetic variants reported to be associated with triglycerides by previous GWAS studies

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	23.77	<i>ASAP3</i>	rs1077514	C	-0.03	6.00E-09
2	169.83	<i>ABCB11</i>	rs2287623	G	0.027	4.00E-12
2	203.53	<i>FAM117B</i>	rs11694172	G	0.028	2.00E-09
2	234.68	<i>UGT1A1</i>	rs11563251	T	0.037	1.00E-09
3	58.38	<i>PXK</i>	rs13315871	A	-0.036	4.00E-08
6	39.25	<i>KCNK17</i>	rs2758886	A	0.023	3.00E-08
6	135.41	<i>HBS1L</i>	rs9376090	C	-0.025	3.00E-09
7	1.08	<i>GPR146</i>	rs1997243	G	0.033	3.00E-10
9	2.64	<i>VLDLR</i>	rs3780181	G	-0.044	7.00E-10
10	17.26	<i>VIM-CUBN</i>	rs10904908	G	0.025	3.00E-11
11	118.49	<i>PHLDB1</i>	rs11603023	T	0.022	1.00E-08
12	9.08	<i>PHC1-A2ML1</i>	rs4883201	G	-0.035	2.00E-09
17	7.09	<i>DLG4</i>	rs314253	C	-0.023	3.00E-10
22	35.71	<i>TOM1</i>	rs138777	A	0.021	5.00E-08
22	46.63	<i>PPARA</i>	rs4253772	T	0.032	1.00E-08

Table 8-14. Novel genetic variants reported to be associated with total cholesterol by previous GWASs

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
7	4411209	<i>SDK1 - SNORD13P2</i>	rs10488360	?	NR	7.00E-06
1	4315204	<i>C1orf174 - AJAP1</i>	rs966321	?	NR	8.00E-06
3	7982827	<i>GRM7 - LMCD1-AS1</i>	rs4591494	?	NR	9.00E-06
13	113760034	<i>F7</i>	rs561241	?	NR	5.00E-16

Table 8-15. Genetic variants reported to be associated with FVII by previous GWASs

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
3	1.85E+08	VPS8	rs4686760	G	0.15	1.00E-06
5	1.12E+08	EPB41L4A	rs379440	G	0.34	1.00E-06
12	90609017	dessert	rs10745527	G	0.16	5.00E-06
19	5582535	SAFB2	rs732505	A	0.25	9.00E-06
2	30526780	LBH	rs6708166	A	0.17	1.00E-06
9	1.37E+08	VAV2	rs12344583	G	0.2	8.00E-06
17	31737421	ACCN1	rs1354492	A	0.16	2.00E-06
6	90310966	ANKRD6	rs6454764	T	0.31	5.00E-06
9	81310680	KRT18P24	rs1757948	G	0.15	7.00E-06
18	25569058	CDH2	rs2298574	G	0.27	6.00E-06
12	1.04E+08	STAB2	rs7306642	A	0.3	3.00E-06

Table 8-16. Genetic variants reported to be associated with FVIII by previous GWASs

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
3	1.85E+08	VPS8	rs4686760	G	0.15	1.00E-06
5	1.12E+08	EPB41L4A	rs379440	G	0.34	1.00E-06
12	90609017	dessert	rs10745527	G	0.16	5.00E-06
19	5582535	SAFB2	rs732505	A	0.25	9.00E-06
2	30526780	LBH	rs6708166	A	0.17	1.00E-06
9	1.37E+08	VAV2	rs12344583	G	0.2	8.00E-06
17	31737421	ACCN1	rs1354492	A	0.16	2.00E-06
6	90310966	ANKRD6	rs6454764	T	0.31	5.00E-06
9	81310680	KRT18P24	rs1757948	G	0.15	7.00E-06
18	25569058	CDH2	rs2298574	G	0.27	6.00E-06
12	1.04E+08	STAB2	rs7306642	A	0.3	3.00E-06

Table 8-17. Genetic variants reported to be associated with vWF by previous GWASs

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
-----	----------	------	----------	-------------	------	---------

1	40064961	<i>PABPC4</i>	rs12037222	A	0.05	6.00E-11
1	66086194	<i>LEPR</i>	rs10889569	T	0.05	9.00E-09
1	66089782	<i>LEPR</i>	rs6700896	T	14.8	3.00E-14
1	66102257	<i>LEPR</i>	rs1805096	A	0.11	2.00E-09
1	66105944	<i>LEPR</i>	rs1892534	A	0.17	7.00E-21
1	66161461	<i>LEPR</i>	rs4420065	C	0.09	4.00E-62
1	154418879	<i>IL6R</i>	rs4537545	T	11.5	2.00E-14
1	154426264	<i>IL6R</i>	rs4129267	C	0.08	2.00E-48
1	154426970	<i>IL6R</i>	rs8192284	?	0.1	2.00E-08
1	159649700	<i>CRP, APCS, DARC, FCER1A, DUSP23, OR10J1, OR10J5, OR10J3, OLFML2B, IFI16, FCRL6</i>	rs16827466	T	0.42	4.00E-73
1	159655518	<i>CRPP1, CRP</i>	rs726640	?	0.44	2.00E-13
1	159674933	<i>CRP</i>	rs876537	C	0.29	1.00E-09
1	159676171	<i>CRP</i>	rs16842559	T	0.11	4.00E-21
1	159678816	<i>CRP</i>	rs2794520	C	0.16	2.00E-186
1	159684665	<i>CRP</i>	rs3091244	?	0.2	6.00E-28
1	159685136	<i>CRP</i>	rs3093059	G	0.16	4.00E-21
1	159692573	<i>CRP, APCS, DARC, FCER1A, DUSP23, OR10J1, OR10J5, OR10J3, OLFML2B, IFI16, FCRL6</i>	rs2808634	T	0.15	3.00E-10
1	159698549	<i>CRP</i>	rs7553007	A	20.7	8.00E-44
1	247601595	<i>NLRP3</i>	rs12239046	C	0.05	1.00E-15
2	27730940	<i>GCKR</i>	rs1260326	T	0.07	5.00E-40
2	27741237	<i>GCKR</i>	rs780094	A	0.14	7.00E-15
2	113841030	<i>IL1F10, IL1RN</i>	rs6734238	G	0.11	9.00E-10
2	113841030	<i>IL1F10</i>	rs6734238	G	0.05	2.00E-17
5	131839618	<i>IRF1</i>	rs4705952	G	0.04	1.00E-08
6	93852252	<i>NR</i>	rs1408282	A	0.41	1.00E-06
6	117114025	<i>GPRC6A</i>	rs6901250	A	0.04	5.00E-08
7	22732839	<i>IL6</i>	rs2097677	A	0.1	4.00E-11
7	72971231	<i>BCL7B</i>	rs13233571	C	0.05	4.00E-09
8	9183358	<i>PPP1R3B</i>	rs9987289	A	0.07	3.00E-13

12	103483094	<i>ASCL1</i>	rs10745954	A	0.04	2.00E-11
12	103495151	<i>Unknown</i>	rs10778213	G	0.12	1.00E-10
12	121402932	<i>HNF1A</i>	rs7305618	T	0.27	1.00E-08
12	121420260	<i>HNF1A</i>	rs7979473	A	0.12	1.00E-10
12	121420807	<i>HNF1A</i>	rs1183910	G	0.15	2.00E-124
12	121424861	<i>HNF1A</i>	rs7310409	A	0.15	7.00E-17
12	121435587	<i>HNF1A, OASL, C12orf43</i>	rs2259816	T	0.14	3.00E-10
12	121439433	<i>HNF1A</i>	rs1169310	A	0.13	2.00E-08
14	73019236	<i>RGS6</i>	rs4903031	G	0.03	5.00E-06
15	60894965	<i>RORA</i>	rs340029	T	0.03	4.00E-09
16	51158710	<i>SALL1</i>	rs10521222	C	0.1	9.00E-13
18	12821593	<i>PTPN2</i>	rs2847281	A	0.03	2.00E-08
19	45395619	<i>APOE</i>	rs2075650	G	0.12	2.00E-21
19	45403412	<i>TOMM40</i>	rs1160985	C	0.13	4.00E-13
19	45410002	<i>APOE</i>	rs769449	?	0.26	9.00E-21
19	45422946	<i>APOC1</i>	rs4420638	A	0.24	9.00E-139
20	43042364	<i>HNF4A</i>	rs1800961	C	0.09	2.00E-09
21	40465534	<i>PSMG1</i>	rs2836878	G	0.03	2.00E-07

Table 8-18. Genetic variants reported to be associated with CRP by previous GWASs.

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	64868352	<i>RNU7-62P</i>	rs11208446	A	0.12	9.00E-06
1	82074852	<i>RP5-837I24.4</i>	rs11809789	A	0.4	6.00E-06
1	159856429	<i>CCDC19</i>	rs2501324	?	0.14	2.00E-10
2	81872922	<i>AC013262.1</i>	rs12052359	A	0.22	7.00E-06
2	88315793	<i>KRCC1</i>	rs12714207	T	0.03	5.00E-07
2	160867059	<i>PLA2R1</i>	rs2667011	?	0.04	2.00E-13
2	190643649	<i>ORMDL1</i>	rs7606224	?	0.1	2.00E-06
2	234264848	<i>DGKD</i>	rs1550532	?	0.04	3.00E-19
2	234406655	<i>USP40</i>	rs6704644	?	0.08	8.00E-08

2	234639310	<i>UGT1A1</i>	rs11891311	?	0.14	1.00E-41
2	234668570	<i>UGT1A1</i>	rs887829	T	0.57	1.00E-69
2	234669144	<i>UGT1A1</i>	rs4148323	?	0.14	3.00E-139
2	234672639	<i>UGT1A1</i>	rs6742078	T	0.23	5E-324
2	234673309	<i>UGT1A</i>	rs4148325	?	0.17	5.00E-62
2	234698790	<i>LOC339766</i>	rs2361502	?	0.1	7.00E-23
3	72023700	<i>RP11-648C16.1</i>	rs2135319	A	0.19	4.00E-06
4	125993502	<i>Intergenic</i>	rs1986655	A	0.02	2.00E-06
4	155389247	<i>DCHS2</i>	rs17031671	?	0.12	6.00E-11
6	26116982	<i>HIST1H2BC</i>	rs12206204	T	0.18	8.00E-07
6	33060822	<i>Intergenic</i>	rs6928954	?	0.08	1.00E-06
7	22756463	<i>AC073072.5</i>	rs10155981	T	0.13	2.00E-07
7	80599357	<i>SEMA3C</i>	rs4236644	A	0.02	2.00E-06
8	445601	<i>C8orf42</i>	rs17665859	?	0.09	5.00E-07
8	112976141	<i>AC068954.1, RP11-58O3.2</i>	rs12549576	G	0.3	1.00E-06
8	137768892	<i>Intergenic</i>	rs16906293	?	0.09	6.00E-06
9	16034538	<i>Intergenic</i>	rs3008706	?	0.06	5.00E-07
9	133765656	<i>Intergenic</i>	rs10901296	?	0.09	3.00E-07
11	2936952	<i>SLC22A18</i>	rs16928809	A	0.06	1.00E-07
12	21017875	<i>SLCO1B3</i>	rs2417940	?	0.06	7.00E-19
12	21074122	<i>SLCO1B3</i>	rs2117032	C	0.13	3.00E-14
12	21331549	<i>SLCO1B1, LST-3TM12, SLCO1A2</i>	rs4149056	C	0.05	7.00E-13
12	21368722	<i>SLCO1B1</i>	rs4363657	?	0.06	5.00E-08
13	70454960	<i>KLHL1</i>	rs11843309	T	0.21	7.00E-06
13	111818832	<i>ARHGEF7</i>	rs4773330	A	0.04	8.00E-06
14	46029687	<i>Intergenic</i>	rs7140958	?	0.09	2.00E-07
16	14045399	<i>ERCC4</i>	rs4781563	A	0.16	2.00E-06
21	34168573	<i>C21orf49</i>	rs2154427	?	0.1	1.00E-06
23	153554404	<i>G6PD</i>	rs766420	G	0.18	9.00E-09

Table 8-19. Genetic variants reported to be associated with Bilirubin by previous GWASs.

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	145723645	<i>PDZK1</i>	rs1967017	T	3.33	4.00E-08
1	145723739	<i>PDZK1</i>	rs1471633	A	0.06	1.00E-29
1	150950062	<i>ARNT</i>	rs4970988	A	0.03	1.00E-07
1	155151493	<i>TRIM46, PKLR</i>	rs11264341	T	0.05	6.00E-19
2	27730940	<i>GCKR</i>	rs1260326	T	0.07	1.00E-44
2	27741237	<i>GCKR</i>	rs780094	T	0.07	5.00E-06
2	27742603	<i>GCKR</i>	rs780093	T	5.15	4.00E-17
2	121306440	<i>INHBB</i>	rs17050272	A	0.04	2.00E-10
2	148716428	<i>ORC4L, ACVR2A</i>	rs2307394	T	0.03	2.00E-08
2	170204846	<i>LRP2</i>	rs2544390	C	0.08	4.00E-08
2	203337001	<i>BMPR2</i>	rs12468226	A	0.04	3.00E-06
3	53100214	<i>SFMBT1, MUSTN1</i>	rs6770152	T	0.04	3.00E-16
4	9915741	<i>SLC2A9</i>	rs11722228	T	0.16	7.00E-24
4	9922167	<i>SLC2A9</i>	rs16890979	T	0.34	7.00E-168
4	9926967	<i>SLC2A9</i>	rs13129697	G	22.21	2.00E-242
4	9934744	<i>SLC2A9</i>	rs737267	C	0.88	3.00E-09
4	9935910	<i>GLUT9</i>	rs6855911	A	0.32	2.00E-16
4	9944052	<i>SLC2A9</i>	rs12498742	A	0.37	1E-700
4	9966380	<i>SLC2A9</i>	rs7442295	C	0.35	3.00E-70
4	9966380	<i>SLC2A9, WDR1</i>	rs7442295	A	0.02	2.00E-15
4	81169912	<i>FGF5</i>	rs11099098	T	0.03	7.00E-07
4	89052323	<i>ABCG2</i>	rs2231142	T	0.22	1.00E-134
4	89054667	<i>ABCG2</i>	rs4148155	G	0.12	1.00E-13
5	39902365	<i>DAB2</i>	rs11954519	A	0.03	4.00E-06
5	55811092	<i>ANKRD55</i>	rs456867	T	0.04	3.00E-06
5	72431482	<i>TMEM171</i>	rs17632159	C	0.04	4.00E-11
6	7102084	<i>RREB1</i>	rs675209	T	0.06	1.00E-23

6	25813150	<i>SLC17A1</i>	rs1165196	G	6.21	5.00E-25
6	25821616	<i>SLC17A1, SLC17A3</i>	rs1165151	T	0.09	7.00E-70
6	25870542	<i>SLC17A3</i>	rs1165205	?	0.09	4.00E-29
6	43272188	<i>SLC22A7</i>	rs4149178	A	0.03	1.00E-06
6	43804571	<i>VEGFA</i>	rs729761	T	0.05	8.00E-16
6	116310287	<i>FRK</i>	rs1933737	T	0.03	7.00E-07
6	134773554	<i>Intergenic</i>	rs9321453	T	0.14	1.00E-09
7	72857049	<i>BAZ1B, MLXIPL</i>	rs1178977	A	0.05	1.00E-12
7	151406005	<i>PRKAG2</i>	rs10480300	T	0.04	4.00E-09
8	23777006	<i>STC1</i>	rs17786744	A	0.03	1.00E-08
8	76478768	<i>HNF4G</i>	rs2941484	T	0.04	4.00E-17
9	33180362	<i>B4GALT1</i>	rs10813960	T	0.03	4.00E-07
10	52646093	<i>A1CF, ASAH2</i>	rs10821905	A	0.06	7.00E-17
10	61469538	<i>SLC16A9</i>	rs1171614	T	0.08	2.00E-28
11	64334114	<i>SLC22A11</i>	rs2078267	T	0.07	9.00E-38
11	64359221	<i>SLC22A12</i>	rs12800450	T	1.19	3.00E-16
11	64440920	<i>SLC22A12</i>	rs506338	C	0.23	2.00E-31
11	64478063	<i>NRXN2, SLC22A12</i>	rs478607	A	0.05	4.00E-11
11	64525216	<i>Intergenic</i>	rs589691	T	0.15	9.00E-08
11	64546391	<i>MEN1, SF1, MAP4K2, PYGM, RASGRP2, CDC42BPG, NRXN2</i>	rs606458	T	0.18	6.00E-11
11	64557054	<i>MAP4K2</i>	rs493573	A	0.8	2.00E-17
11	65560620	<i>OVOL1, LTBP3</i>	rs642803	T	0.04	3.00E-13
11	119235404	<i>USP2</i>	rs2195525	T	0.03	2.00E-06
12	48173352	<i>HDAC7</i>	rs4760636	T	0.03	6.00E-06
12	52251272	<i>ACVR1B, ACVRL1</i>	rs7976059	T	0.03	2.00E-09
12	57809456	<i>R3HDM2, INHBC</i>	rs1106766	T	5.16	2.00E-11
12	57844049	<i>INHBC, INHBE</i>	rs3741414	T	0.07	2.00E-25
12	112007756	<i>ATXN2, PTPN11</i>	rs653178	T	0.04	7.00E-12
12	121416988	<i>HNF1A</i>	rs2244608	A	0.04	8.00E-08

12	122625992	<i>B3GNT4</i>	rs7953704	A	0.03	3.00E-08
13	72345505	<i>DACH1</i>	rs584480	T	0.02	9.00E-06
15	73082366	<i>ADPGK</i>	rs4777542	T	0.03	2.00E-07
15	76158983	<i>UBE2Q2, NRG4</i>	rs1394125	A	0.04	3.00E-13
15	99271135	<i>IGF1R</i>	rs6598541	A	0.04	5.00E-15
16	69563890	<i>NFAT5</i>	rs7193778	T	0.05	8.00E-10
16	79734987	<i>MAF</i>	rs7188445	A	0.03	2.00E-09
17	53364788	<i>HLF</i>	rs7224610	A	0.04	5.00E-17
17	59465697	<i>BCAS3, C17orf82</i>	rs2079742	T	0.04	1.00E-08
17	74283669	<i>QRICH2, PRPSAP1</i>	rs164009	A	0.03	2.00E-07
18	57872989	<i>MC4R</i>	rs12955983	A	0.04	2.00E-06
19	7199803	<i>INSR</i>	rs1035942	A	0.03	1.00E-07

Table 8-20. Genetic variants reported to be associated with urate by previous GWASs.

Chr	Position	Gene	lead SNP	Risk Allele	beta	p-Value
1	152300817	<i>FLG</i>	rs3126085	A	1.22	6.00E-12
1	152442289	<i>FLG</i>	rs6661961	T	0.34	9.00E-11
2	102971865	<i>IL1RL1, IL18R1, IL18RAP</i>	rs13015714	G	1.27	8.00E-18
2	112003867	<i>LOC100505634</i>	rs2271404	G	1.17	3.00E-07
3	33065339	<i>GLB1</i>	rs7613051	A	1.29	6.00E-21
3	112376308	<i>CCDC80, LOC100630917</i>	rs12634229	G	1.29	2.00E-19
5	131995843	<i>IL13, RAD50</i>	rs1295686	A	1.38	2.00E-17
5	132049027	<i>KIF3A</i>	rs2897442	C	1.11	4.00E-08
6	31272321	<i>HLA-C</i>	rs9368677	G	1.36	1.00E-17
6	31499603	<i>BAT1</i>	rs3853601	G	1.13	2.00E-06
6	32074804	<i>CREBL1, TNXB,</i>	rs12153855	T	1.58	3.00E-14
6	32158319	<i>GPSM3</i>	rs176095	T	1.4	8.00E-20
6	32308908	<i>C6orf10</i>	rs9469099	G	1.61	5.00E-19
7	3128789	<i>CARD11</i>	rs4722404	G	1.18	8.00E-09
8	81308150	<i>ZBTB10</i>	rs7000782	A	1.09	1.00E-06

8	129427518	MIR1208	rs7815944	A	1.16	4.00E-07
10	64398466	ZNF365	rs10995251	C	1.28	6.00E-20
10	64560470	ADO, EGR2	rs1444418	A	1.17	2.00E-07
11	7968359	OR10A3, NLRP10	rs878860	G	1.31	2.00E-22
11	65513107	OVOL1	rs593982	C	1.23	6.00E-07
11	65551957	OVOL1	rs479844	G	1.14	1.00E-13
11	76270683	C11orf30, LRRC32	rs7130588	G	1.29	4.00E-13
11	76301316	C11orf30	rs7927894	A	1.22	8.00E-10
11	76332210	C11orf30	rs11236809	G	1.24	3.00E-06
16	11210415	CLEC16A	rs9923856	A	1.17	6.00E-06
19	8789381	ACTL9	rs2164983	A	1.16	7.00E-09
20	52807221	CYP24A1, PFDN4	rs16999165	T	1.19	2.00E-08
22	37258503	NCF4	rs4821544	C	1.09	6.00E-06

Table 8-21. Genetic variants reported to be associated with atopic dermatitis by previous GWASs.

SNP	A1	A2	freq	b	se	p	N
rs6720173	C	G	0.1583	0.2561	0.0386	3.23E-11	15213
rs10208987	C	A	0.0799	0.3782	0.0502	4.77E-14	15213
rs4148189	A	G	0.1139	0.2538	0.0441	8.66E-09	15213
rs4245786	G	A	0.2425	0.1423	0.0333	1.88E-05	15213
rs4299376	C	A	0.3154	-0.2829	0.0325	3.10E-18	15213
rs4953023	A	G	0.0658	0.7625	0.0523	3.57E-48	15213

Table 8-22. SNPs significantly associated with GBD in the GBD meta-analysis

Marker ID	A1	A2	Freq	beta	StdErr	P.value	N	CHR	traits	Gene
rs10429888	A	G	0.3558	4.0061	0.6687	2.09E-09	7105	1	QTC interval	<i>OLFML2B</i> (dist=31343), <i>NOS1AP</i> (dist=14594)
rs10918594	C	G	0.5689	2.8147	0.4769	3.59E-09	7101	1	QTC interval	<i>OLFML2B</i> (dist=37044), <i>NOS1AP</i> (dist=8893)
rs12143842	T	C	0.4025	3.726	0.517	5.70E-13	7097	1	QTC interval	<i>OLFML2B</i> (dist=40246), <i>NOS1AP</i> (dist=5691)
rs16857031	C	G	0.6309	2.8037	0.6321	9.19E-06	7104	1	QTC interval	<i>NOS1AP</i>
rs10918740	T	C	0.4521	2.78	0.4663	2.49E-09	7105	1	QTC interval	<i>NOS1AP</i>
rs12039600	A	G	0.3555	4.2885	0.709	1.46E-09	7106	1	QTC interval	<i>NOS1AP</i>
rs4657166	C	G	0.5729	2.3324	0.4936	2.30E-06	7105	1	QTC interval	<i>NOS1AP</i>
rs72633699	T	C	0.3724	3.6318	0.6375	1.22E-08	7105	1	QTC interval	<i>NOS1AP</i>
rs10918859	A	G	0.3783	3.5549	0.5865	1.35E-09	7106	1	QTC interval	<i>NOS1AP</i>
rs4657178	T	C	0.4091	2.5549	0.5192	8.61E-07	7103	1	QTC interval	<i>NOS1AP</i>
rs10919024	A	C	0.6366	3.3784	0.6591	2.97E-07	7105	1	QTC interval	<i>NOS1AP</i>
rs7645178	A	G	0.5828	1.9978	0.4512	9.51E-06	6902	3	PR interval	<i>SCN5A</i>
rs41312411	C	G	0.6545	2.8977	0.5877	8.18E-07	6904	3	PR interval	<i>SCN5A</i>
rs12635898	A	C	0.3909	2.5793	0.4753	5.73E-08	6807	3	PR interval	<i>SCN5A</i>
rs62242769	A	G	0.3687	-2.3914	0.5032	2.01E-06	6902	3	PR interval	<i>SCN5A</i>
rs55824920	T	C	0.4144	2.4975	0.4449	1.97E-08	6898	3	PR interval	<i>SCN5A</i>
rs6795580	C	G	0.4609	-2.271	0.4298	1.27E-07	6902	3	PR interval	<i>SCN5A</i>
rs9818148	T	G	0.6118	-2.4408	0.4792	3.52E-07	6883	3	PR interval	<i>SCN5A</i>
rs41312045	C	G	0.2839	5.4286	1.179	4.14E-06	6903	3	PR interval	<i>SCN5A</i>
rs73056438	T	C	0.3572	2.6866	0.5318	4.38E-07	6902	3	PR interval	<i>SCN5A</i>
rs6599230	T	C	0.3604	-2.6812	0.5272	3.66E-07	6894	3	PR interval	<i>SCN5A</i>
rs6773331	A	T	0.1946	-8.1823	1.5691	1.84E-07	6903	3	PR interval	<i>SCN5A</i>
rs4131768	A	G	0.5796	-2.0936	0.4482	2.99E-06	6904	3	PR interval	<i>SCN5A</i> (dist=4010), <i>SCN10A</i> (dist=43663)
rs182475051	T	C	0.388	-2.6333	0.4812	4.46E-08	6902	3	PR interval	<i>SCN5A</i> (dist=34328), <i>SCN10A</i> (dist=13345)
rs11129800	T	C	0.4677	2.9887	0.4149	5.86E-13	6902	3	PR interval	<i>SCN10A</i>
rs9990137	A	G	0.5727	-2.4132	0.441	4.45E-08	6902	3	PR interval	<i>SCN10A</i>

rs73062575	T	G	0.2642	5.2862	1.1925	9.29E-06	6900	3	PR interval	SCN10A
rs6801957	T	C	0.4526	3.9874	0.4243	5.61E-21	6903	3	PR interval	SCN10A
rs10428168	T	C	0.5829	-2.499	0.4411	1.46E-08	6902	3	PR interval	SCN10A
rs12636123	T	G	0.528	-2.5996	0.423	7.98E-10	6902	3	PR interval	SCN10A
rs7553796	A	C	0.4426	-0.0635	0.0093	8.14E-12	9528	1	il6	IL6R
rs7518199	A	C	0.5524	0.0883	0.0092	9.49E-22	9533	1	il6	IL6R
rs4072391	T	C	0.2896	-0.0581	0.0119	1.03E-06	9540	1	il6	IL6R
rs9998811	A	G	0.4517	-0.0226	0.0023	5.53E-23	4242	4	sUrate	SLC2A9
rs7696092	A	C	0.5487	-0.0203	0.0023	3.40E-19	4241	4	sUrate	SLC2A9
rs2741029	T	G	0.5653	2.0051	0.0955	5.84E-98	6521	2	bilirubin	UGT1A8
rs7608713	A	G	0.4166	-0.9284	0.1042	5.11E-19	6530	2	bilirubin	UGT1A10,UGT1A8
rs17868322	A	G	0.3576	1.603	0.2346	8.36E-12	6526	2	bilirubin	UGT1A10,UGT1A8
rs28899170	A	C	0.4362	2.4878	0.0933	9.74E-157	6522	2	bilirubin	UGT1A10,UGT1A6,UGT1A7,UGT1A8,UGT1A9
rs1604144	T	C	0.4403	-1.1735	0.0988	1.58E-32	6493	2	bilirubin	UGT1A10,UGT1A6,UGT1A7,UGT1A8,UGT1A9
rs17862875	A	G	0.4291	2.5304	0.0942	5.68E-159	6530	2	bilirubin	UGT1A10,UGT1A3,UGT1A4,UGT1A5,UGT1A6,UGT1A7,UGT1A8,UGT1A9
rs11568318	A	C	0.3609	-1.5878	0.1859	1.31E-17	6501	2	bilirubin	UGT1A10,UGT1A3,UGT1A4,UGT1A5,UGT1A6,UGT1A7,UGT1A8,UGT1A9
rs11563251	T	C	0.3683	-1.237	0.1435	6.64E-18	6518	2	bilirubin	UGT1A1,UGT1A10,UGT1A3,UGT1A4,UGT1A5,UGT1A6,UGT1A7,UGT1A8,UGT1A9
rs7586006	T	G	0.6176	-0.7347	0.1316	2.39E-08	6529	2	bilirubin	UGT1A10(dist=3576),HJURP(dist=59959)
rs10209214	T	C	0.554	-0.9548	0.0953	1.24E-23	6523	2	bilirubin	UGT1A10(dist=5632),HJURP(dist=57903)
rs1115381	T	C	0.4221	-0.9004	0.1025	1.53E-18	6483	2	bilirubin	USP40(dist=35418),UGT1A8(dist=21006)
rs2741027	A	G	0.4346	2.0055	0.0954	3.60E-98	6530	2	bilirubin	USP40(dist=48144),UGT1A8(dist=8280)
rs491098	C	G	0.199	-21.5146	0.5994	3.64E-282	8737	13	FVII	F7
rs3093253	A	G	0.1906	-22.3575	0.6249	2.33E-280	8739	13	FVII	F7
rs563964	T	C	0.4867	4.9929	0.4071	1.42E-34	8735	13	FVII	F10
rs547138	A	T	0.5802	3.9481	0.4233	1.09E-20	8733	13	FVII	F10
rs867186	A	G	0.8115	7.3879	0.724	1.89E-24	8738	20	FVII	PROCR
rs17406518	A	C	0.8102	5.5107	0.7059	5.85E-15	8737	20	FVII	PROCR(dist=34011),MMP24(dist=15280)

rs1058003	A	G	0.5948	2.1242	0.4231	5.16E-07	8739	20	FVII	<i>TRPC4AP</i>
rs6060230	T	C	0.187	7.3503	0.7349	1.50E-23	8738	20	FVII	<i>TRPC4AP(dist=8690),EDEM2(dist=13852)</i>
rs11906160	A	G	0.1953	5.4717	0.6851	1.38E-15	8738	20	FVII	<i>MYH7B</i>
rs7261167	A	G	0.184	7.4098	0.7441	2.32E-23	8738	20	FVII	<i>MYH7B</i>
rs8176725	A	G	0.3073	11.5058	1.1053	2.25E-25	5867	9	FVIII	<i>ABO</i>
rs2073827	C	G	0.4723	-7.761	0.645	2.42E-33	5868	9	FVIII	<i>ABO</i>
rs8176694	T	C	0.6578	-5.455	0.8213	3.10E-11	5861	9	FVIII	<i>ABO</i>
rs8176644	T	C	0.2984	16.206	1.2834	1.49E-36	5871	9	FVIII	<i>ABO</i>
rs507666	A	G	0.3517	14.3829	0.791	7.07E-74	5871	9	FVIII	<i>ABO</i>
rs630014	A	G	0.4934	-8.2094	0.6295	7.19E-39	5870	9	FVIII	<i>ABO</i>
rs11244061	T	C	0.3287	13.2379	0.9874	5.47E-41	5872	9	FVIII	<i>ABO(dist=3351),SURF6(dist=43571)</i>
rs633862	T	C	0.5037	-7.4347	0.6316	5.49E-32	5864	9	FVIII	<i>ABO(dist=4814),SURF6(dist=42108)</i>
rs554710	T	C	0.6048	-5.4182	0.6924	5.09E-15	5869	9	FVIII	<i>ABO(dist=31218),SURF6(dist=15704)</i>
rs1179011	A	G	0.5708	4.8883	0.6587	1.16E-13	5860	9	FVIII	<i>ABO(dist=35753),SURF6(dist=11169)</i>
rs11244084	T	C	0.3075	9.818	1.2502	4.07E-15	5871	9	FVIII	<i>ABO(dist=40380),SURF6(dist=6542)</i>
rs10901262	A	G	0.3857	4.6887	0.7122	4.60E-11	5870	9	FVIII	<i>ABO(dist=41486),SURF6(dist=5436)</i>
rs11244034	T	C	0.3318	4.119	0.8678	2.07E-06	5870	9	FVIII	<i>OBP2B</i>
rs12554449	T	G	0.5365	5.6209	1.2358	5.41E-06	2871	9	FVIII	<i>OBP2B</i>
rs3761823	C	G	0.3494	3.679	0.7977	3.99E-06	5864	9	FVIII	<i>OBP2B(dist=2184),ABO(dist=43751)</i>
rs185852900	A	G	0.3202	7.1986	0.9549	4.77E-14	5869	9	FVIII	<i>OBP2B(dist=18035),ABO(dist=27900)</i>
rs2013075	T	C	0.2941	13.6835	1.2919	3.25E-26	5859	9	FVIII	<i>OBP2B(dist=18525),ABO(dist=27410)</i>
rs182267720	T	C	0.3206	7.2223	0.9875	2.59E-13	5869	9	FVIII	<i>OBP2B(dist=29716),ABO(dist=16219)</i>
rs9411470	A	G	0.663	4.552	0.8636	1.36E-07	5772	9	FVIII	<i>OBP2B(dist=35357),ABO(dist=10578)</i>
rs10901250	A	G	0.3184	6.848	0.965	1.28E-12	5869	9	FVIII	<i>OBP2B(dist=36998),ABO(dist=8937)</i>
rs11244049	T	C	0.7053	15.9482	1.2859	2.54E-35	5855	9	FVIII	<i>OBP2B(dist=38464),ABO(dist=7471)</i>
rs7857390	A	G	0.4615	-7.5723	0.6468	1.17E-31	5872	9	FVIII	<i>OBP2B(dist=43918),ABO(dist=2017)</i>
rs10901253	T	C	0.6255	-5.3952	0.7465	4.94E-13	5871	9	FVIII	<i>OBP2B(dist=44144),ABO(dist=1791)</i>

rs28715750	T	C	0.7242	-6.9293	1.5568	8.55E-06	5871	9	FVIII	<i>SURF4</i>
rs3118660	A	T	0.3031	-5.8465	1.0295	1.36E-08	5871	9	FVIII	<i>SURF4</i>
rs3758348	C	G	0.3395	11.0928	0.9249	3.82E-33	5869	9	FVIII	<i>SURF4</i>
rs62575992	T	C	0.3169	-4.8132	0.9725	7.44E-07	5871	9	FVIII	<i>C9orf96</i>
rs77926044	T	C	0.2999	-5.4358	1.1644	3.04E-06	5871	9	FVIII	<i>C9orf96</i>
rs17474001	T	C	0.3357	10.7113	0.9439	7.65E-30	5870	9	FVIII	<i>C9orf96</i>
rs3124747	A	G	0.4066	3.2712	0.6822	1.62E-06	5872	9	FVIII	<i>C9orf96</i>
rs41302673	T	G	0.6826	10.0792	1.1259	3.48E-19	5872	9	FVIII	<i>C9orf96</i>
rs36219252	A	G	0.3108	13.2123	2.4447	6.50E-08	5872	9	FVIII	<i>ADAMTS13</i>
rs72779205	A	G	0.3266	-5.5328	0.9186	1.71E-09	5868	9	FVIII	<i>ADAMTS13</i>
rs3118671	T	G	0.3394	-5.567	0.8268	1.66E-11	5867	9	FVIII	<i>ADAMTS13</i>
rs41296094	T	C	0.7142	10.6838	1.5532	6.04E-12	5872	9	FVIII	<i>ADAMTS13</i>
rs28645493	C	G	0.683	9.6009	1.1685	2.10E-16	5869	9	FVIII	<i>ADAMTS13</i>
rs652600	A	G	0.6	-3.8721	0.6977	2.87E-08	5869	9	FVIII	<i>ADAMTS13</i>
rs2769073	T	C	0.3292	-3.9742	0.8797	6.25E-06	5871	9	FVIII	<i>ADAMTS13</i>
rs2769074	T	C	0.3017	-5.1076	1.0782	2.17E-06	5872	9	FVIII	<i>ADAMTS13</i>
rs4962153	A	G	0.3336	9.4932	0.9352	3.28E-24	5872	9	FVIII	<i>ADAMTS13</i>
rs3124764	T	C	0.3352	9.4455	0.9361	6.10E-24	5872	9	FVIII	<i>CACFD1</i>
rs41302667	A	G	0.3114	9.1834	1.1719	4.65E-15	5869	9	FVIII	<i>CACFD1</i>
rs2073936	T	C	0.2165	15.3003	3.3628	5.37E-06	5871	9	FVIII	<i>CACFD1</i>
rs3094381	A	G	0.3029	-5.0661	1.0744	2.42E-06	5872	9	FVIII	<i>CACFD1</i>
rs35434910	A	T	0.3639	5.0703	0.7664	3.71E-11	5869	9	FVIII	<i>GBGT1(dist=3023),OBP2B(dist=38342)</i>
rs7022409	T	C	0.4266	-3.5256	0.6606	9.47E-08	5871	9	FVIII	<i>GBGT1(dist=3969),OBP2B(dist=37396)</i>
rs6597604	A	G	0.5998	3.5731	0.6914	2.36E-07	5870	9	FVIII	<i>GBGT1(dist=10479),OBP2B(dist=30886)</i>
rs186686586	T	C	0.363	-3.5115	0.7691	4.98E-06	5869	9	FVIII	<i>GBGT1(dist=21428),OBP2B(dist=19937)</i>
rs12378537	T	C	0.2985	10.7287	1.1452	7.37E-21	5872	9	FVIII	<i>GBGT1(dist=26225),OBP2B(dist=15140)</i>
rs11244031	A	G	0.471	-4.1714	0.6391	6.71E-11	5865	9	FVIII	<i>GBGT1(dist=36863),OBP2B(dist=4502)</i>

rs4783961	A	G	0.5096	-0.0309	0.0041	2.74E-14	14310	16	HDL	<i>CETP</i>
rs1800775	A	C	0.4856	0.0682	0.004	5.99E-64	14312	16	HDL	<i>CETP</i>
rs1864163	A	G	0.3016	-0.0729	0.0045	8.18E-58	14312	16	HDL	<i>CETP</i>
rs9929488	C	G	0.3016	-0.0549	0.0046	4.52E-33	14311	16	HDL	<i>CETP</i>
rs118146573	A	G	0.1848	-0.081	0.0059	2.73E-42	14324	16	HDL	<i>CETP</i>
rs11076174	T	C	0.8517	-0.0496	0.0075	2.99E-11	14326	16	HDL	<i>CETP</i>
rs7205804	A	G	0.4449	0.0692	0.0041	2.33E-64	14315	16	HDL	<i>CETP</i>
rs5883	T	C	0.1264	0.0566	0.0092	7.49E-10	13528	16	HDL	<i>CETP</i>
rs289714	A	G	0.7721	-0.0628	0.0053	1.51E-32	14315	16	HDL	<i>CETP</i>
rs289715	A	T	0.1796	0.0383	0.0063	1.06E-09	14319	16	HDL	<i>CETP</i>
rs289719	T	C	0.3415	0.0442	0.0044	5.22E-24	14309	16	HDL	<i>CETP</i>
rs5880	C	G	0.1224	-0.0794	0.0088	3.01E-19	14325	16	HDL	<i>CETP</i>
rs2562126	A	G	0.1946	-0.0307	0.0058	1.52E-07	14324	16	HDL	<i>HERPUD1</i>
rs9938160	T	C	0.6614	0.0223	0.0044	4.54E-07	14321	16	HDL	<i>HERPUD1(dist=6797),CETP(dist=11245)</i>
rs9989419	A	G	0.4105	-0.0498	0.0042	4.67E-33	14312	16	HDL	<i>HERPUD1(dist=7346),CETP(dist=10696)</i>
rs193695	A	G	0.3749	-0.0482	0.0043	1.10E-29	14313	16	HDL	<i>HERPUD1(dist=7363),CETP(dist=10679)</i>
rs72786786	A	G	0.3375	0.0671	0.0044	2.36E-52	14308	16	HDL	<i>HERPUD1(dist=7721),CETP(dist=10321)</i>
rs12448528	A	G	0.2635	-0.0581	0.0051	3.09E-30	12079	16	HDL	<i>HERPUD1(dist=7762),CETP(dist=10280)</i>
rs7205692	A	G	0.7897	-0.035	0.0056	3.37E-10	14323	16	HDL	<i>HERPUD1(dist=9121),CETP(dist=8921)</i>
rs12920974	T	G	0.3171	-0.0479	0.0045	3.18E-26	14318	16	HDL	<i>HERPUD1(dist=15232),CETP(dist=2810)</i>
rs12708967	T	C	0.7623	-0.0617	0.0051	1.87E-33	14324	16	HDL	<i>HERPUD1(dist=15418),CETP(dist=2624)</i>
rs3764261	A	C	0.3502	0.0822	0.0043	7.35E-80	14317	16	HDL	<i>HERPUD1(dist=15531),CETP(dist=2511)</i>
rs1077834	T	C	0.7407	0.049	0.0049	2.54E-23	14322	15	HDL	<i>LIPC</i>
rs8033940	A	G	0.3197	0.0381	0.0045	4.29E-17	14289	15	HDL	<i>LIPC</i>
rs12914035	A	G	0.7965	-0.0299	0.0056	1.08E-07	14295	15	HDL	<i>LIPC</i>
rs261338	A	G	0.2205	0.0517	0.0055	3.35E-21	14323	15	HDL	<i>LIPC</i>
rs261336	A	G	0.7849	0.0463	0.0055	4.41E-17	14325	15	HDL	<i>LIPC</i>

rs74304285	A	G	0.1991	0.0384	0.0057	2.36E-011	14323	8	HDL	<i>LPL</i>
rs343	A	C	0.1441	0.0503	0.0072	3.50E-012	14320	8	HDL	<i>LPL</i>
rs269	T	G	0.7683	0.0395	0.0052	4.19E-014	14322	8	HDL	<i>LPL</i>
rs301	T	C	0.7228	0.0396	0.0048	1.36E-016	14324	8	HDL	<i>LPL</i>
rs314	A	G	0.3161	0.0315	0.0045	2.13E-012	14326	8	HDL	<i>LPL</i>
rs13702	T	C	0.6773	0.0399	0.0045	3.98E-019	14324	8	HDL	<i>LPL</i>
rs1569209	T	G	0.8564	0.0638	0.0071	2.30E-019	14319	8	HDL	<i>LPL(dist=5400),SLC18A1(dist=172196)</i>
rs10096633	T	C	0.1833	0.059	0.006	7.51E-023	14326	8	HDL	<i>LPL(dist=6151),SLC18A1(dist=171445)</i>
rs35237252	A	C	0.2903	0.0397	0.0047	2.32E-017	14323	8	HDL	<i>LPL(dist=45501),SLC18A1(dist=132095)</i>
rs4333617	A	G	0.552	0.029	0.0041	2.23E-012	14323	8	HDL	<i>LPL(dist=50234),SLC18A1(dist=127362)</i>
rs191343955	T	C	0.6845	0.0302	0.0045	1.88E-011	14279	8	HDL	<i>LPL(dist=90514),SLC18A1(dist=87082)</i>
rs9644568	A	G	0.1878	0.0489	0.006	2.74E-016	14324	8	HDL	<i>LPL(dist=103812),SLC18A1(dist=73784)</i>
rs7015766	T	C	0.1348	0.054	0.0077	2.44E-012	14311	8	HDL	<i>LPL(dist=114279),SLC18A1(dist=63317)</i>
rs405509	T	G	0.4909	0.088	0.0119	1.34E-13	14056	19	LDL	<i>APOE</i>
rs769449	A	G	0.2098	0.2097	0.0179	9.61E-32	14048	19	LDL	<i>APOE</i>
rs7412	T	C	0.1762	-0.484	0.022	9.00E-107	13226	19	LDL	<i>APOE</i>
rs445925	A	G	0.1951	-0.3159	0.0192	1.10E-60	14057	19	LDL	<i>APOE(dist=2990),APOC1(dist=2281)</i>
rs6859	A	G	0.4382	0.0725	0.0119	1.22E-09	14044	19	LDL	<i>PVRL2</i>
rs283813	A	T	0.1673	-0.1163	0.0234	6.69E-07	14007	19	LDL	<i>PVRL2</i>
rs7254892	A	G	0.1303	-0.4105	0.0337	3.84E-34	14057	19	LDL	<i>PVRL2</i>
rs6857	T	C	0.2414	0.2042	0.0164	1.37E-35	13154	19	LDL	<i>PVRL2</i>
rs157580	A	G	0.5781	-0.0952	0.0124	1.32E-14	13264	19	LDL	<i>TOMM40</i>
rs2075650	A	G	0.7757	0.2018	0.0168	3.40E-33	14057	19	LDL	<i>TOMM40</i>
rs61679753	A	T	0.2384	-0.3502	0.0345	2.98E-24	14057	19	LDL	<i>TOMM40</i>
rs115881343	T	C	0.1332	0.1703	0.0369	3.96E-06	14047	19	LDL	<i>TOMM40</i>
rs6511720	T	G	0.202	-0.1922	0.0178	3.59E-27	14056	19	LDL	<i>LDLR</i>
rs1010679	T	C	0.7191	-0.0933	0.0142	4.99E-11	14054	19	LDL	<i>LDLR</i>

rs2228671	T	C	0.2086	-0.1489	0.0176	2.50E-17	14053	19	LDL	<i>LDLR</i>
rs2738447	A	C	0.4291	-0.0608	0.0123	7.68E-07	13175	19	LDL	<i>LDLR</i>
rs5742911	A	G	0.6383	-0.069	0.0137	4.99E-07	12044	19	LDL	<i>LDLR</i>
rs1529729	T	C	0.5331	0.0562	0.0119	2.13E-06	14056	19	LDL	<i>SMARCA4</i>
rs3786725	A	G	0.3492	-0.0577	0.0127	5.91E-06	14055	19	LDL	<i>SMARCA4</i>
rs73015007	A	G	0.3076	-0.0814	0.0136	1.89E-09	14055	19	LDL	<i>SMARCA4(dist=10879),LDLR(dist=16201)</i>
rs11668477	A	G	0.7264	-0.1145	0.0144	1.58E-15	14057	19	LDL	<i>SMARCA4(dist=22072),LDLR(dist=5008)</i>
rs17248720	T	C	0.2004	-0.1904	0.0179	1.76E-26	14056	19	LDL	<i>SMARCA4(dist=25229),LDLR(dist=1851)</i>
rs10858082	A	G	0.5594	-0.0662	0.012	3.94E-08	14048	1	LDL	<i>CELSR2</i>
rs4970834	T	C	0.2572	-0.1294	0.0151	1.16E-17	14055	1	LDL	<i>CELSR2</i>
rs7528419	A	G	0.7177	-0.1502	0.0143	7.73E-26	14014	1	LDL	<i>CELSR2</i>
rs11102967	T	C	0.6371	-0.1127	0.0127	5.66E-19	14043	1	LDL	<i>CELSR2</i>
rs11577931	A	G	0.8197	-0.1384	0.0203	9.29E-12	14054	1	LDL	<i>CELSR2(dist=2506),PSRC1(dist=1292)</i>
rs1042031	T	C	0.2499	-0.0934	0.0155	1.55E-09	14057	2	LDL	<i>APOB</i>
rs693	A	G	0.5209	-0.096	0.0119	6.18E-16	14057	2	LDL	<i>APOB</i>
rs113549125	T	G	0.4476	0.101	0.012	3.17E-17	14052	2	LDL	<i>APOB</i>
rs12713956	A	G	0.8264	-0.0978	0.0207	2.43E-06	14056	2	LDL	<i>APOB</i>
rs41288783	A	G	0.2504	0.9996	0.2069	1.36E-06	7498	2	LDL	<i>APOB</i>
rs12691202	T	C	0.1406	-0.143	0.0299	1.68E-06	14057	2	LDL	<i>APOB</i>
rs1367117	A	G	0.369	0.1163	0.0125	1.15E-20	14055	2	LDL	<i>APOB</i>
rs17398765	A	G	0.8217	0.1077	0.0223	1.33E-06	14057	2	LDL	<i>APOB(dist=3806),LOC645949(dist=639555)</i>
rs1713222	A	G	0.2278	-0.1629	0.0166	7.52E-23	14056	2	LDL	<i>APOB(dist=4378),LOC645949(dist=638983)</i>
rs71435594	T	C	0.173	-0.1832	0.0223	2.34E-16	14055	2	LDL	<i>APOB(dist=42390),LOC645949(dist=600971)</i>
rs111526071	T	C	0.3312	0.0966	0.0131	1.53E-13	14048	2	LDL	<i>APOB(dist=46834),LOC645949(dist=596527)</i>
rs97458	A	G	0.6588	-0.0947	0.0129	1.76E-13	14057	2	LDL	<i>APOB(dist=60690),LOC645949(dist=582671)</i>
rs4635554	T	G	0.6305	0.0755	0.0126	2.07E-09	13975	2	LDL	<i>APOB(dist=122714),LOC645949(dist=520647)</i>
rs507627	T	C	0.3019	-0.1048	0.0137	1.83E-14	14057	2	LDL	<i>APOB(dist=136884),LOC645949(dist=506477)</i>

rs365946	T	G	0.2383	-0.1226	0.0157	6.44E-15	14055	2	LDL	<i>APOB(dist=172384),LOC645949(dist=470977)</i>
rs11900894	A	C	0.704	0.0824	0.0139	2.91E-09	14024	2	LDL	<i>APOB(dist=183319),LOC645949(dist=460042)</i>
rs436566	T	C	0.2483	-0.1014	0.0156	7.91E-11	14056	2	LDL	<i>APOB(dist=184659),LOC645949(dist=458702)</i>
rs35239705	A	G	0.2626	-0.0955	0.0147	9.29E-11	14054	2	LDL	<i>APOB(dist=184742),LOC645949(dist=458619)</i>
rs7571647	A	G	0.1921	-0.1497	0.0191	3.96E-15	14056	2	LDL	<i>APOB(dist=209689),LOC645949(dist=433672)</i>
rs769449	A	G	0.245	-0.3076	0.0224	4.34E-43	11537	19	crp	<i>APOE</i>
rs75627662	T	C	0.2961	-0.1641	0.0184	5.00E-19	11543	19	crp	<i>APOE</i>
rs439401	T	C	0.4055	0.0986	0.0154	1.36E-10	11544	19	crp	<i>APOE(dist=1801),APOC1(dist=3470)</i>
rs1205	T	C	0.3857	-0.1662	0.0156	1.60E-26	11538	1	crp	<i>CRP</i>
rs3091244	A	G	0.4039	0.1865	0.0152	1.33E-34	11539	1	crp	<i>CRP</i>
rs11579148	T	C	0.1826	-0.3538	0.0632	2.16E-08	11545	1	crp	<i>CRP(dist=35023),DUSP23(dist=31357)</i>
rs12815613	A	G	0.6248	0.115	0.0157	2.35E-13	11542	12	crp	<i>HNF1A-AS1</i>
rs2244608	A	G	0.6288	-0.1649	0.0157	1.20E-25	11543	12	crp	<i>HNF1A</i>
rs1169286	T	C	0.5511	-0.1237	0.0148	5.79E-17	11541	12	crp	<i>HNF1A</i>
rs2245407	A	C	0.197	-0.1345	0.0293	4.47E-06	11543	12	crp	<i>HNF1A</i>
rs12427353	C	G	0.2851	0.1176	0.0186	2.46E-10	11541	12	crp	<i>HNF1A</i>
rs1169300	A	G	0.3559	-0.1562	0.0161	2.58E-22	11543	12	crp	<i>HNF1A</i>
rs1169303	A	C	0.504	-0.1204	0.0146	1.98E-16	11541	12	crp	<i>HNF1A</i>
rs1169307	T	C	0.411	0.1146	0.0151	2.70E-14	11540	12	crp	<i>HNF1A</i>
rs3751152	C	G	0.3672	0.0981	0.0158	5.29E-10	11515	12	crp	<i>C12orf43</i>
rs3751151	A	T	0.4398	0.0924	0.0155	2.72E-09	11533	12	crp	<i>C12orf43</i>
rs73214128	T	C	0.187	-0.1428	0.0317	6.71E-06	11542	12	crp	<i>C12orf43</i>
rs1169314	A	G	0.6426	-0.1581	0.016	6.83E-23	11528	12	crp	<i>C12orf43</i>
rs2264779	T	C	0.5901	0.1157	0.0151	1.66E-14	11538	12	crp	<i>C12orf43</i>
rs6589566	A	G	0.7801	0.1362	0.0127	7.91E-27	13636	11	TG	<i>ZNF259</i>
rs35120633	A	G	0.2184	0.1387	0.0129	7.02E-27	13636	11	TG	<i>ZNF259</i>
rs3741298	T	C	0.6958	0.0962	0.0079	9.36E-34	13631	11	TG	<i>ZNF259</i>

rs12287066	T	G	0.2187	0.1359	0.0128	1.74E-26	13636	11	TG	<i>APOA5</i>
rs662799	A	G	0.789	0.1444	0.0134	4.18E-27	13635	11	TG	<i>APOA5</i>
rs6589567	A	C	0.2551	0.0836	0.0101	8.67E-17	13635	11	TG	<i>APOA5(dist=7540),APOA4(dist=20742)</i>
rs61905108	A	C	0.205	0.1322	0.0149	6.09E-19	13579	11	TG	<i>BUD13</i>
rs180326	T	G	0.5978	0.0499	0.0066	4.87E-14	13608	11	TG	<i>BUD13</i>
rs10488699	T	C	0.2205	0.1311	0.0125	1.19E-25	13633	11	TG	<i>BUD13</i>
rs17440824	T	C	0.2581	0.0504	0.0096	1.52E-07	13634	11	TG	<i>BUD13</i>
rs10790162	A	G	0.2194	0.1356	0.0127	1.23E-26	13636	11	TG	<i>BUD13</i>
rs1263151	T	C	0.5069	0.0344	0.0063	4.00E-08	13631	11	TG	<i>BUD13</i>
rs2041967	A	G	0.7046	0.0385	0.0082	2.52E-06	13635	11	TG	<i>BUD13(dist=1435),ZNF259(dist=4127)</i>
rs499790	T	C	0.2473	0.0766	0.0104	1.50E-13	13628	11	TG	<i>LOC283143(dist=888821),BUD13(dist=99147)</i>
rs481843	T	C	0.2293	0.09	0.0118	2.32E-14	13610	11	TG	<i>LOC283143(dist=894949),BUD13(dist=93019)</i>
rs486394	A	C	0.6314	0.0405	0.0069	4.31E-09	13632	11	TG	<i>LOC283143(dist=895404),BUD13(dist=92564)</i>
rs11820504	T	C	0.71	0.0442	0.0083	8.80E-08	13614	11	TG	<i>LOC283143(dist=898524),BUD13(dist=89444)</i>
rs509728	C	G	0.3288	0.0775	0.0164	2.34E-06	13635	11	TG	<i>LOC283143(dist=902249),BUD13(dist=85719)</i>
rs112064887	A	T	0.4339	0.0317	0.0064	6.01E-07	13357	11	TG	<i>LOC283143(dist=903370),BUD13(dist=84598)</i>
rs61906105	A	G	0.7955	0.1128	0.0144	4.09E-15	13635	11	TG	<i>LOC283143(dist=919610),BUD13(dist=68358)</i>
rs12417015	A	G	0.822	0.1098	0.0216	3.79E-07	13636	11	TG	<i>LOC283143(dist=919858),BUD13(dist=68110)</i>
rs12799766	A	G	0.3275	0.0442	0.0074	2.25E-09	13633	11	TG	<i>LOC283143(dist=927509),BUD13(dist=60459)</i>
rs1145210	T	C	0.3881	-0.0337	0.0067	5.48E-07	13630	11	TG	<i>LOC283143(dist=928084),BUD13(dist=59884)</i>
rs11216103	A	G	0.2102	0.1027	0.0128	1.16E-15	13636	11	TG	<i>LOC283143(dist=943570),BUD13(dist=44398)</i>
rs12221682	C	G	0.6467	0.0507	0.0112	6.44E-06	13622	11	TG	<i>LOC283143(dist=947299),BUD13(dist=40669)</i>
rs11216107	T	C	0.624	0.0439	0.0068	1.04E-10	13636	11	TG	<i>LOC283143(dist=948745),BUD13(dist=39223)</i>
rs7350481	T	C	0.212	0.1193	0.0133	2.62E-19	13636	11	TG	<i>LOC283143(dist=955365),BUD13(dist=32603)</i>
rs75273015	A	G	0.1983	0.1556	0.021	1.36E-13	13632	11	TG	<i>LOC283143(dist=962912),BUD13(dist=25056)</i>
rs180360	A	G	0.6525	-0.0374	0.0072	1.81E-07	13603	11	TG	<i>LOC283143(dist=968070),BUD13(dist=19898)</i>
rs12279180	T	C	0.7757	0.1175	0.012	1.46E-22	13636	11	TG	<i>LOC283143(dist=973114),BUD13(dist=14854)</i>

rs1558860	A	C	0.2171	0.1298	0.0127	1.16E-24	13636	11	TG	LOC283143(dist=976450),BUD13(dist=11518)
rs61905084	T	C	0.6954	-0.0494	0.0086	1.13E-08	12101	11	TG	LOC283143(dist=979376),BUD13(dist=8592)
rs57232565	T	C	0.2301	0.1178	0.0142	1.19E-16	12828	11	TG	LOC283143(dist=980539),BUD13(dist=7429)
rs10466533	A	G	0.7703	0.1125	0.0119	2.27E-21	13613	11	TG	LOC283143(dist=981210),BUD13(dist=6758)
rs56321069	A	T	0.3342	-0.0394	0.0073	7.09E-08	13630	8	TG	LPL
rs74304285	A	G	0.2715	-0.0549	0.0089	7.71E-10	13633	8	TG	LPL
rs343	A	C	0.2252	-0.0734	0.0113	9.07E-11	13632	8	TG	LPL
rs253	T	C	0.4634	-0.0332	0.0063	1.22E-07	13634	8	TG	LPL
rs269	T	G	0.7065	-0.0547	0.0081	1.92E-11	13633	8	TG	LPL
rs285	T	C	0.4789	-0.0447	0.0063	1.44E-12	13564	8	TG	LPL
rs314	A	G	0.3594	-0.0561	0.0069	5.50E-16	13636	8	TG	LPL
rs10096633	T	C	0.2527	-0.0939	0.0094	1.39E-23	13636	8	TG	LPL(dist=6151),SLC18A1(dist=171445)
rs117604010	A	G	0.2032	-0.1255	0.0243	2.41E-07	13636	8	TG	LPL(dist=23626),SLC18A1(dist=153970)
rs17411024	A	G	0.2556	-0.0638	0.0094	1.37E-11	13636	8	TG	LPL(dist=27364),SLC18A1(dist=150232)
rs894210	A	G	0.5217	0.0477	0.0063	3.01E-14	13634	8	TG	LPL(dist=41073),SLC18A1(dist=136523)
rs35237252	A	C	0.3402	-0.074	0.0072	1.26E-24	13633	8	TG	LPL(dist=45501),SLC18A1(dist=132095)
rs7005359	A	G	0.7006	-0.0444	0.0081	3.51E-08	13458	8	TG	LPL(dist=45923),SLC18A1(dist=131673)
rs4333617	A	G	0.5393	-0.0448	0.0064	1.74E-12	13634	8	TG	LPL(dist=50234),SLC18A1(dist=127362)
rs10103634	A	G	0.3849	-0.0381	0.0067	1.36E-08	13631	8	TG	LPL(dist=65842),SLC18A1(dist=111754)
rs6586891	A	C	0.5952	0.0454	0.0067	9.34E-12	13214	8	TG	LPL(dist=89828),SLC18A1(dist=87768)
rs9644568	A	G	0.2589	-0.0759	0.0093	3.46E-16	13634	8	TG	LPL(dist=103812),SLC18A1(dist=73784)
rs11985911	T	C	0.4096	0.0416	0.0066	2.36E-10	13634	8	TG	LPL(dist=107318),SLC18A1(dist=70278)
rs28597716	A	G	0.6981	-0.0452	0.008	1.61E-08	13636	8	TG	LPL(dist=111917),SLC18A1(dist=65679)
rs7015766	T	C	0.2168	-0.0769	0.012	1.37E-10	13622	8	TG	LPL(dist=114279),SLC18A1(dist=63317)
rs80336612	T	C	0.5686	0.0334	0.0065	2.33E-07	13635	8	TG	LPL(dist=115452),SLC18A1(dist=62144)
rs13265868	A	G	0.4573	-0.0292	0.0064	4.52E-06	13636	8	TG	LPL(dist=118257),SLC18A1(dist=59339)
rs8176725	A	G	0.2909	0.127	0.0088	2.03E-047	9004	9	vWF	ABO

rs512770	A	G	0.3427	-0.0661	0.0062	1.98E-026	9007	9	vWF	ABO
rs2073826	T	G	0.4653	-0.0982	0.005	4.34E-085	9008	9	vWF	ABO
rs8176694	T	C	0.6722	-0.0694	0.0065	6.58E-027	8993	9	vWF	ABO
rs8176765	A	C	0.774	0.1583	0.0166	1.35E-021	9000	9	vWF	ABO
rs8176632	T	C	0.3112	0.0714	0.0073	1.47E-022	9007	9	vWF	ABO(dist=1917),SURF6(dist=45005)
rs8176630	T	C	0.3462	-0.0683	0.0061	7.66E-029	9003	9	vWF	ABO(dist=2092),SURF6(dist=44830)
rs11244061	T	C	0.3074	0.1714	0.0077	1.07E-110	9009	9	vWF	ABO(dist=3351),SURF6(dist=43571)
rs558240	A	G	0.4419	0.0731	0.005	7.12E-049	9006	9	vWF	ABO(dist=6503),SURF6(dist=40419)
rs554710	T	C	0.6121	-0.0555	0.0055	5.55E-024	9006	9	vWF	ABO(dist=31218),SURF6(dist=15704)
rs75444660	A	G	0.7149	0.1815	0.0112	3.90E-059	9009	9	vWF	ABO(dist=31845),SURF6(dist=15077)
rs1179011	A	G	0.5773	0.0577	0.0052	2.54E-028	8986	9	vWF	ABO(dist=35753),SURF6(dist=11169)
rs11244084	T	C	0.2852	0.1423	0.0098	7.65E-048	9008	9	vWF	ABO(dist=40380),SURF6(dist=6542)
rs12379461	A	G	0.4368	0.0246	0.0054	5.50E-06	8249	9	vWF	OBP2B
rs12554449	T	G	0.4139	-0.0488	0.0083	3.54E-09	5934	9	vWF	OBP2B
rs10793957	T	G	0.679	0.0422	0.0068	5.39E-10	9008	9	vWF	OBP2B(dist=3303),ABO(dist=42632)
rs185852900	A	G	0.3037	0.073	0.0075	1.78E-22	9003	9	vWF	OBP2B(dist=18035),ABO(dist=27900)
rs2013075	T	C	0.2747	0.1616	0.0102	5.78E-57	8995	9	vWF	OBP2B(dist=18525),ABO(dist=27410)
rs138683771	A	C	0.7267	0.1357	0.0193	1.90E-12	9007	9	vWF	OBP2B(dist=20935),ABO(dist=25000)
rs182267720	T	C	0.3025	0.0839	0.0078	3.36E-27	9005	9	vWF	OBP2B(dist=29716),ABO(dist=16219)
rs11793768	A	G	0.2786	-0.0439	0.0086	3.59E-07	9009	9	vWF	OBP2B(dist=32034),ABO(dist=13901)
rs7855466	T	C	0.31	0.0755	0.0073	5.82E-25	9001	9	vWF	OBP2B(dist=36675),ABO(dist=9260)
rs7025839	A	G	0.2864	-0.0412	0.0087	1.88E-06	9008	9	vWF	OBP2B(dist=39562),ABO(dist=6373)
rs7857390	A	G	0.4551	-0.0945	0.0051	1.03E-77	9009	9	vWF	OBP2B(dist=43918),ABO(dist=2017)
rs77693339	T	C	0.7194	0.1838	0.0101	2.85E-74	9006	9	vWF	OBP2B(dist=43930),ABO(dist=2005)
rs12554339	A	C	0.6381	-0.076	0.0059	1.58E-38	9006	9	vWF	OBP2B(dist=44109),ABO(dist=1826)
rs1179034	T	C	0.7184	-0.05	0.0088	1.62E-08	9006	9	vWF	SURF4
rs3118660	A	T	0.285	-0.0589	0.0081	2.71E-13	9008	9	vWF	SURF4

rs28406995	A	G	0.2858	-0.0397	0.0089	7.41E-06	9008	9	vWF	<i>SURF4</i>
rs56343119	A	C	0.3191	0.1338	0.0072	2.41E-77	9009	9	vWF	<i>SURF4</i>
rs34399743	A	G	0.2879	-0.04	0.0087	4.64E-06	9008	9	vWF	<i>C9orf96</i>
rs62575992	T	C	0.3044	-0.0601	0.0078	9.39E-15	9008	9	vWF	<i>C9orf96</i>
rs77926044	T	C	0.2796	-0.0545	0.0091	2.33E-09	9008	9	vWF	<i>C9orf96</i>
rs17474001	T	C	0.3159	0.1284	0.0073	1.42E-68	9007	9	vWF	<i>C9orf96</i>
rs3124747	A	G	0.403	0.04	0.0054	8.99E-14	9008	9	vWF	<i>C9orf96</i>
rs190371202	T	G	0.6915	-0.0364	0.0074	7.95E-07	9008	9	vWF	<i>C9orf96</i>
rs41302673	T	G	0.7046	0.138	0.0088	1.18E-55	9009	9	vWF	<i>C9orf96</i>
rs34265876	T	C	0.7107	-0.0394	0.0088	6.99E-06	9009	9	vWF	<i>ADAMTS13</i>
rs34024143	T	C	0.3046	-0.0415	0.0074	2.07E-08	9009	9	vWF	<i>ADAMTS13</i>
rs36219252	A	G	0.2832	0.1803	0.0192	5.03E-21	9009	9	vWF	<i>ADAMTS13</i>
rs72779205	A	G	0.3121	-0.0425	0.0073	6.38E-09	9005	9	vWF	<i>ADAMTS13</i>
rs3118671	T	G	0.3272	-0.0565	0.0065	5.52E-18	9002	9	vWF	<i>ADAMTS13</i>
rs41296094	T	C	0.7272	0.106	0.0125	1.91E-17	9009	9	vWF	<i>ADAMTS13</i>
rs28645493	C	G	0.7061	0.1328	0.0091	3.09E-48	9006	9	vWF	<i>ADAMTS13</i>
rs592514	A	T	0.7019	-0.0448	0.0077	5.69E-09	9008	9	vWF	<i>ADAMTS13</i>
rs652600	A	G	0.6075	-0.042	0.0055	2.30E-14	9005	9	vWF	<i>ADAMTS13</i>
rs2769073	T	C	0.3147	-0.0459	0.007	4.81E-11	9007	9	vWF	<i>ADAMTS13</i>
rs4962153	A	G	0.3155	0.1248	0.0073	8.98E-66	9009	9	vWF	<i>ADAMTS13</i>
rs739468	T	G	0.3141	0.1241	0.0073	2.83E-65	9008	9	vWF	<i>CACFD1</i>
rs78534112	A	G	0.2369	0.1044	0.0155	1.59E-11	9001	9	vWF	<i>CACFD1</i>
rs41302667	A	G	0.2895	0.1276	0.0091	3.12E-44	9006	9	vWF	<i>CACFD1</i>
rs2073936	T	C	0.2176	0.1434	0.0279	2.62E-07	9008	9	vWF	<i>CACFD1</i>
rs3094381	A	G	0.2863	-0.0455	0.0085	8.40E-08	9009	9	vWF	<i>CACFD1</i>
rs621907	T	C	0.276	-0.0497	0.0095	1.48E-07	9009	9	vWF	<i>MED22</i>
rs512050	T	C	0.2782	-0.0484	0.0093	2.08E-07	9009	9	vWF	<i>MED22</i>

rs117119759	A	G	0.2708	0.1688	0.0142	1.81E-32	9009	9	vWF	MED22
rs41307428	T	C	0.287	0.1811	0.0192	3.99E-21	9008	9	vWF	SLC2A6
rs3094378	A	G	0.2828	-0.0425	0.0087	1.07E-06	9009	9	vWF	SLC2A6
rs3124762	A	C	0.6735	-0.0442	0.0067	3.05E-11	9006	9	vWF	SLC2A6
rs41297217	A	G	0.2982	-0.0553	0.0086	1.29E-10	8228	9	vWF	SLC2A6
rs41309954	A	C	0.2654	-0.0623	0.0097	1.63E-10	9005	9	vWF	SLC2A6
rs3124760	T	C	0.6824	-0.0421	0.0072	3.94E-09	9005	9	vWF	SLC2A6
rs2073935	T	G	0.3823	0.0437	0.0056	8.94E-15	9002	9	vWF	SLC2A6
rs3094326	C	G	0.3361	-0.0512	0.0066	7.70E-15	9008	9	vWF	SLC2A6
rs78602133	T	C	0.2816	-0.0447	0.0098	5.06E-06	9009	9	vWF	SLC2A6(dist=3244),TMEM8C(dist=32188)
rs72779222	T	C	0.7424	-0.0502	0.0106	2.19E-06	9007	9	vWF	SLC2A6(dist=3556),TMEM8C(dist=31876)
rs4259477	C	G	0.341	-0.0375	0.0063	3.18E-09	8982	9	vWF	SLC2A6(dist=4809),TMEM8C(dist=30623)
rs3094324	T	C	0.3763	0.0289	0.0056	2.92E-07	8998	9	vWF	SLC2A6(dist=5874),TMEM8C(dist=29558)
rs2157780	C	G	0.6126	-0.0298	0.0055	7.73E-08	9006	9	vWF	SLC2A6(dist=8845),TMEM8C(dist=26587)
rs183925556	A	G	0.3568	0.0299	0.006	7.16E-07	9007	9	vWF	SLC2A6(dist=9592),TMEM8C(dist=25840)
rs77771223	T	C	0.2989	-0.0454	0.0082	3.14E-08	9008	9	vWF	SLC2A6(dist=12170),TMEM8C(dist=23262)
rs3094325	A	G	0.2792	0.0797	0.0136	4.63E-09	8989	9	vWF	SLC2A6(dist=16601),TMEM8C(dist=18831)
rs186264304	A	G	0.6815	0.0805	0.0072	1.00E-28	9009	9	vWF	SLC2A6(dist=18859),TMEM8C(dist=16573)
rs28615587	T	G	0.3962	0.0288	0.0054	1.01E-07	9009	9	vWF	SLC2A6(dist=20870),TMEM8C(dist=14562)
rs35434910	A	T	0.3536	0.0588	0.006	1.90E-22	9006	9	vWF	GBGT1(dist=3023),OBP2B(dist=38342)
rs34187118	T	C	0.2956	-0.0512	0.0081	2.04E-10	9009	9	vWF	GBGT1(dist=3150),OBP2B(dist=38215)
rs7022409	T	C	0.4214	-0.044	0.0052	3.11E-17	9008	9	vWF	GBGT1(dist=3969),OBP2B(dist=37396)
rs4284147	T	C	0.3549	-0.0377	0.006	2.91E-10	9007	9	vWF	GBGT1(dist=7838),OBP2B(dist=33527)
rs35333961	A	G	0.4014	0.0413	0.0054	2.67E-14	9006	9	vWF	GBGT1(dist=16864),OBP2B(dist=24501)
rs7028259	T	C	0.6627	-0.0336	0.0064	1.65E-07	9007	9	vWF	GBGT1(dist=22347),OBP2B(dist=19018)
rs12378537	T	C	0.2828	0.1144	0.0091	1.97E-36	9009	9	vWF	GBGT1(dist=26225),OBP2B(dist=15140)
rs7039497	A	G	0.2845	-0.0497	0.0087	1.04E-08	9008	9	vWF	GBGT1(dist=30630),OBP2B(dist=10735)

rs11244031	A	G	0.4642	-0.0481	0.0051	1.88E-21	9002	9	vWF	<i>GBGT1(dist=36863),OBP2B(dist=4502)</i>
rs10901243	C	G	0.4289	0.0275	0.0052	1.22E-07	8998	9	vWF	<i>GBGT1(dist=39074),OBP2B(dist=2291)</i>
rs405509	T	G	0.4903	0.0605	0.0132	4.58E-06	14594	19	TC	<i>APOE</i>
rs769449	A	G	0.2102	0.2101	0.0198	2.87E-26	14586	19	TC	<i>APOE</i>
rs7412	T	C	0.1763	-0.3898	0.0246	2.38E-56	13758	19	TC	<i>APOE</i>
rs439401	T	C	0.3931	-0.0605	0.0137	9.87E-06	14594	19	TC	<i>APOE(dist=1801),APOC1(dist=3470)</i>
rs445925	A	G	0.1949	-0.2479	0.0214	4.71E-31	14595	19	TC	<i>APOE(dist=2990),APOC1(dist=2281)</i>
rs6511720	T	G	0.2019	-0.2254	0.0199	7.35E-30	14594	19	TC	<i>LDLR</i>
rs6511721	A	G	0.5154	0.0636	0.0131	1.13E-06	14591	19	TC	<i>LDLR</i>
rs73015030	A	G	0.1544	-0.1665	0.0349	1.91E-06	14595	19	TC	<i>LDLR</i>
rs2228671	T	C	0.208	-0.1759	0.0196	2.89E-19	14591	19	TC	<i>LDLR</i>
rs2738447	A	C	0.4287	-0.0664	0.0137	1.31E-06	13704	19	TC	<i>LDLR</i>
rs5742911	A	G	0.6387	-0.0735	0.0153	1.52E-06	12541	19	TC	<i>LDLR</i>
rs8099996	A	G	0.6059	-0.0646	0.0135	1.82E-06	14590	19	TC	<i>SMARCA4(dist=1667),LDLR(dist=25413)</i>
rs8102273	T	C	0.6486	-0.0696	0.0141	7.98E-07	14591	19	TC	<i>SMARCA4(dist=7089),LDLR(dist=19991)</i>
rs73015007	A	G	0.3066	-0.0974	0.0151	1.03E-10	14593	19	TC	<i>SMARCA4(dist=10879),LDLR(dist=16201)</i>
rs11668477	A	G	0.7273	-0.1431	0.016	3.08E-19	14595	19	TC	<i>SMARCA4(dist=22072),LDLR(dist=5008)</i>
rs17248720	T	C	0.2001	-0.2253	0.0199	1.29E-29	14594	19	TC	<i>SMARCA4(dist=25229),LDLR(dist=1851)</i>
rs7254892	A	G	0.131	-0.3109	0.0375	1.05E-16	14595	19	TC	<i>PVRL2</i>
rs6857	T	C	0.2435	0.1949	0.0182	1.05E-26	13684	19	TC	<i>PVRL2</i>
rs1042031	T	C	0.2505	-0.0849	0.0172	7.75E-07	14595	2	TC	<i>APOB</i>
rs693	A	G	0.5211	-0.1044	0.0132	2.22E-15	14595	2	TC	<i>APOB</i>
rs2854725	T	G	0.8361	-0.1384	0.0254	5.03E-08	14588	2	TC	<i>APOB</i>
rs41288783	A	G	0.2021	1.0301	0.2088	8.06E-07	8671	2	TC	<i>APOB</i>
rs10199768	T	G	0.4616	0.1065	0.0133	9.36E-16	14592	2	TC	<i>APOB</i>
rs12714102	C	G	0.168	-0.1499	0.0237	2.61E-10	14592	2	TC	<i>APOB</i>
rs1367117	A	G	0.3691	0.1195	0.0139	6.62E-18	14593	2	TC	<i>APOB</i>

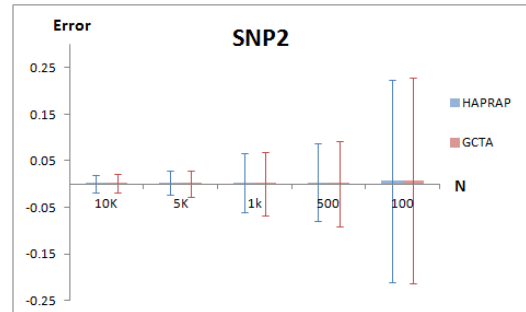
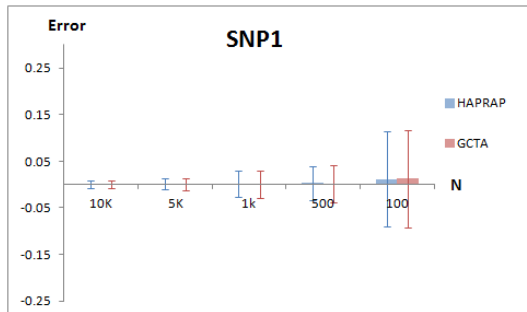
rs72653053	T	C	0.8791	-0.2604	0.0574	5.80E-06	14595	2	TC	APOB
rs1800481	A	G	0.228	-0.1779	0.0184	5.15E-22	14552	2	TC	APOB
rs17398765	A	G	0.8211	0.1133	0.0247	4.55E-06	14595	2	TC	APOB(dist=3806),LOC645949(dist=639555)
rs541041	A	G	0.7509	-0.1635	0.017	7.42E-22	14593	2	TC	APOB(dist=28030),LOC645949(dist=615331)
rs61539677	A	G	0.6288	0.11	0.0139	3.09E-15	14589	2	TC	APOB(dist=38250),LOC645949(dist=605111)
rs71435594	T	C	0.1739	-0.1728	0.0249	3.67E-12	14593	2	TC	APOB(dist=42390),LOC645949(dist=600971)
rs97458	A	G	0.6576	-0.1047	0.0143	2.48E-13	14595	2	TC	APOB(dist=60690),LOC645949(dist=582671)
rs12477249	A	G	0.3721	0.087	0.0139	3.49E-10	14591	2	TC	APOB(dist=133716),LOC645949(dist=509645)
rs507627	T	C	0.302	-0.1082	0.0152	1.09E-12	14595	2	TC	APOB(dist=136884),LOC645949(dist=506477)
rs11903224	C	G	0.2968	0.0973	0.0154	2.58E-10	14533	2	TC	APOB(dist=164416),LOC645949(dist=478945)
rs380240	T	C	0.2386	-0.1425	0.0175	3.41E-16	14592	2	TC	APOB(dist=179708),LOC645949(dist=463653)
rs436566	T	C	0.2488	-0.0911	0.0173	1.47E-07	14594	2	TC	APOB(dist=184659),LOC645949(dist=458702)
rs35239705	A	G	0.2632	-0.1021	0.0164	5.11E-10	14592	2	TC	APOB(dist=184742),LOC645949(dist=458619)
rs386397	A	G	0.3535	0.0735	0.014	1.60E-07	14591	2	TC	APOB(dist=184882),LOC645949(dist=458479)
rs10198972	A	G	0.1477	-0.1453	0.0328	9.42E-06	14595	2	TC	APOB(dist=186266),LOC645949(dist=457095)
rs312049	T	C	0.5705	0.0609	0.0132	4.33E-06	14592	2	TC	APOB(dist=208999),LOC645949(dist=434362)
rs7571647	A	G	0.1925	-0.1547	0.0212	3.05E-13	14594	2	TC	APOB(dist=209689),LOC645949(dist=433672)

Table 8-23. 394 SNPs selected for fine mapping analysis in the UCLEB meta-analysis of CVD related traits

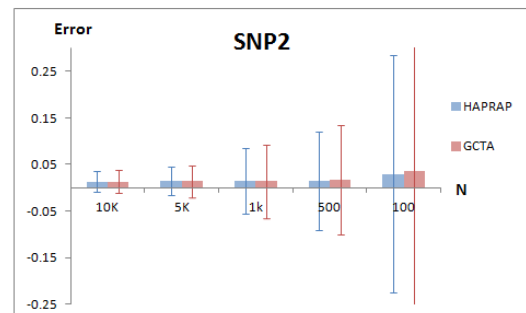
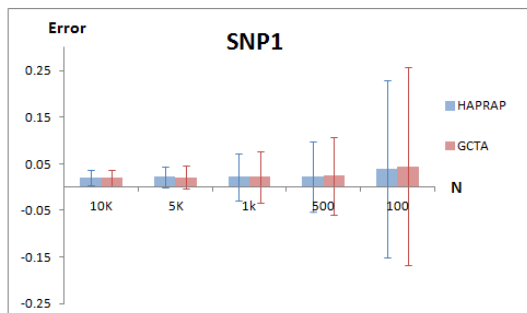
8.3. Supplementary Figures

A

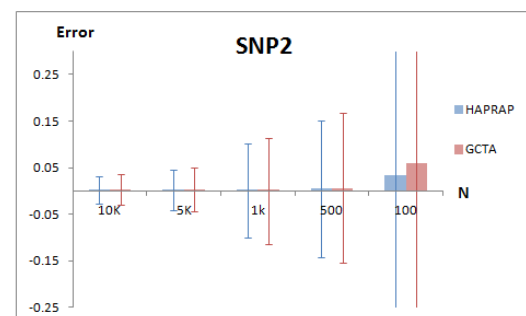
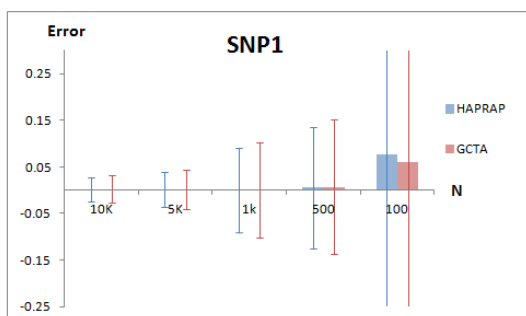
$r^2=0.2$



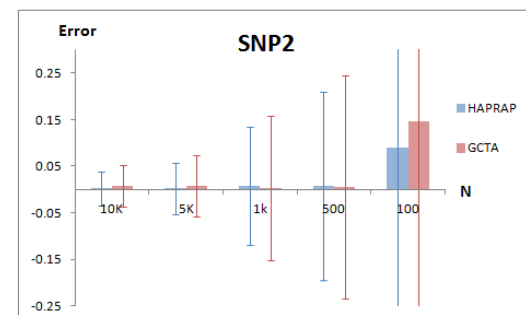
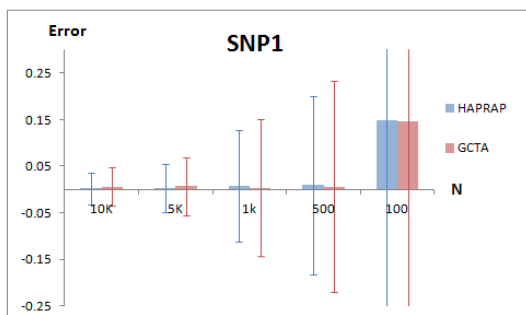
$r^2=0.5$



$r^2=0.8$

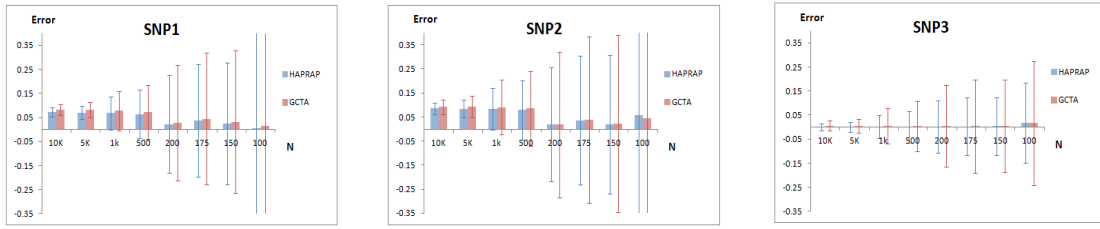


$r^2=0.9$

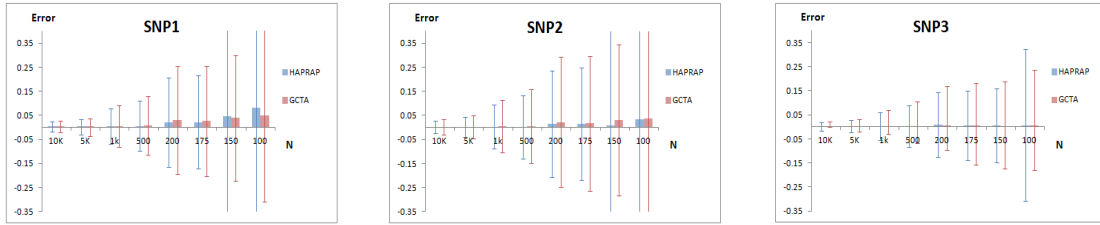


B

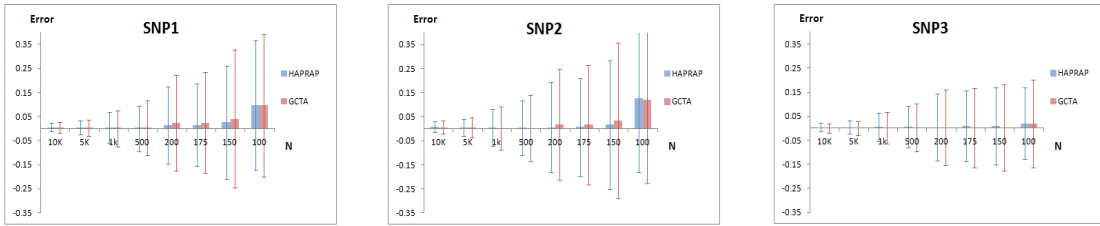
$r^2=0.8$ between SNP1 & SNP2 / $r^2=0.5$ between SNP1 & SNP3



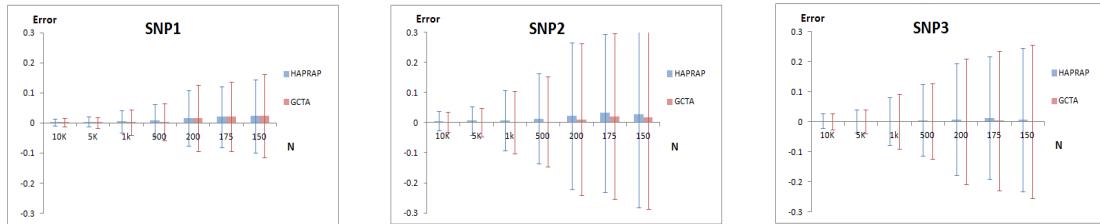
$r^2=0.8$ between SNP1 & SNP2 / $r^2=0.3$ between SNP1 & SNP3



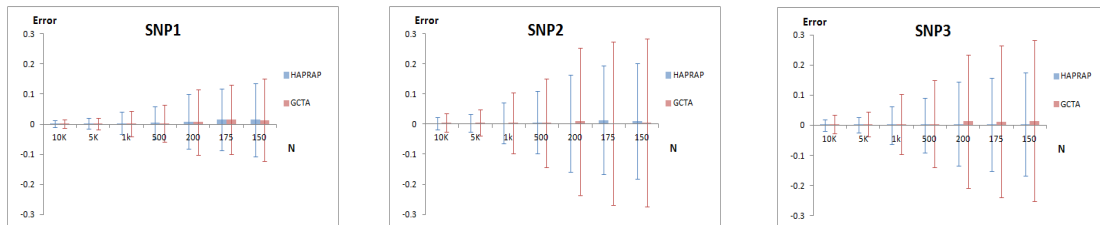
$r^2=0.8$ between SNP1 & SNP2 / $r^2=0.1$ between SNP1 & SNP3



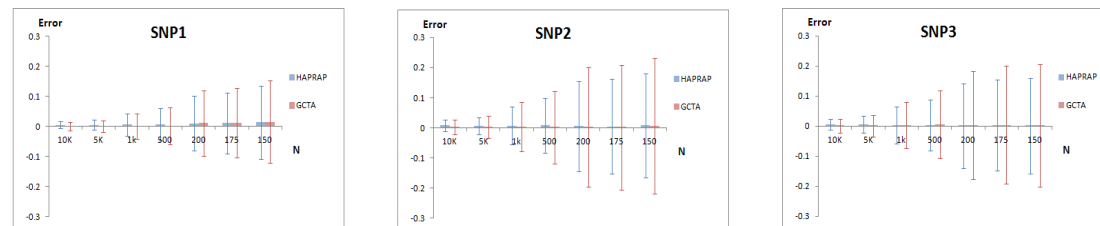
$r^2=0.5$ between SNP1 & SNP2 / $r^2=0.5$ between SNP1 & SNP3



$r^2=0.5$ between SNP1 & SNP2 / $r^2=0.3$ between SNP1 & SNP3

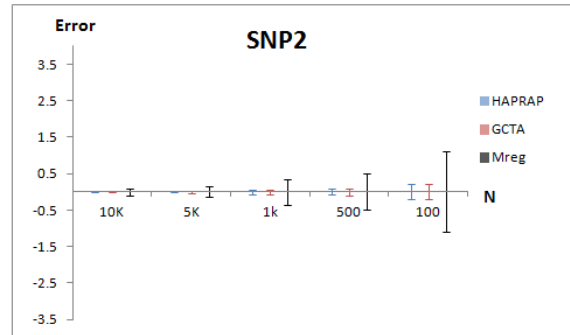
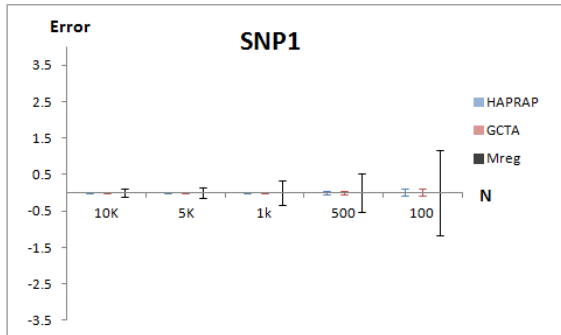


$r^2=0.5$ between SNP1 & SNP2 / $r^2=0.1$ between SNP1 & SNP3

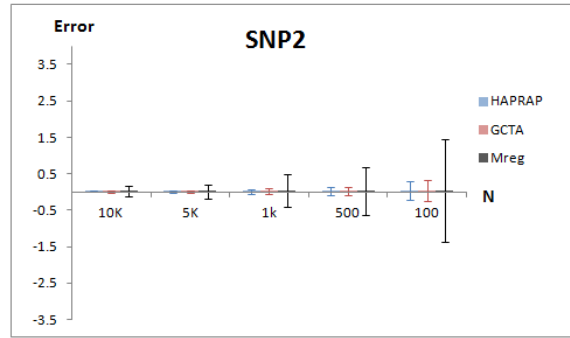
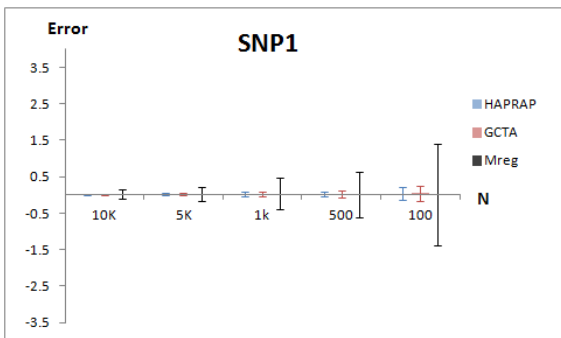


C

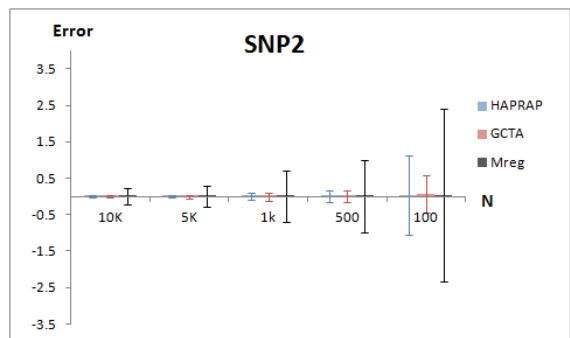
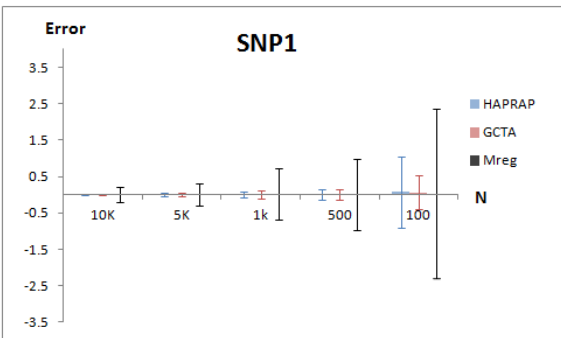
$r^2=0.2$



$r^2=0.5$



$r^2=0.8$



$r^2=0.9$

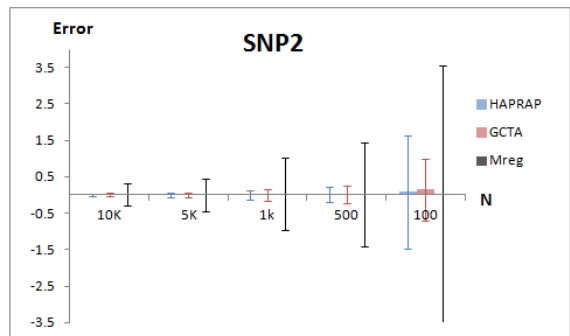
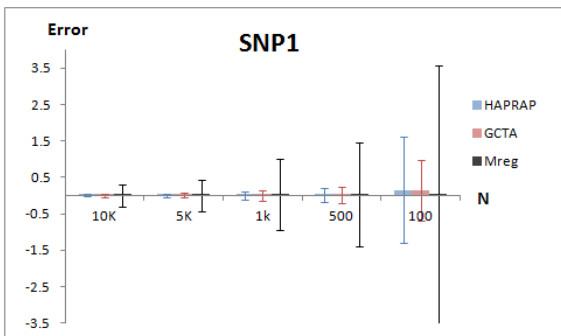


Figure 8-1. *Performance comparison of HAPRAP, GCTA and multiple regression using artificial meta-analyses of the simulated populations.* Error (represented as a bar) is defined as the difference between the gold standard and the mean (with 95% CI) of 1000 replications. N is the number of individuals in the reference panel. **A**, the comparison against HAPRAP and GCTA for the 2-SNPs models with $r^2=0.8$. **B**, the comparison against HAPRAP and GCTA for the 3-SNPs models. **C**, the comparison among HAPRAP, GCTA using meta-analysis data and multiple regression using individual-level data from the reference panel.